

A unified political trajectory

Since reunification, Germany has witnessed significant scientific development almost regardless of the party in charge. The forthcoming election offers smooth continuation and important challenges, whoever wins.

As German citizens prepare to go to the polls, it is worth reflecting on the state of German research when the country was reunified 12 years ago by Chancellor Helmut Kohl, head of the long-standing Christian Democrat–Free Democrat coalition. The inward-looking East German Academy of Sciences employed tens of thousands of researchers, all secure in their jobs but producing little research that could be considered internationally competitive. The comfortable academic research organizations in the west employed hundreds of thousands of researchers, most of them secure in their jobs, and producing very good, internationally competitive research — although not enough to have justified their generous budgets.

The task of closing the vast gap in science policy began. The academy was duly dissolved, and huge investment was made in research structures in the east which were newly created in the image of those in western Germany. The widely acknowledged stodginess of the west also had to be tackled, to convert the over-bureaucratized, over-hierarchical, over-regulated research environment into something much more responsive and flexible. The *Zeitgeist* dictated also that closer links between academia and industry be forged.

Much of this has been achieved, and independently of party lines. Kohl's coalition, which enjoyed two legislative periods after reunification, started the process. Its successor, Gerhard Schröder's Social Democrat–Green coalition, continued it. The next government — and at the moment it is anyone's guess what its political shade will be — will almost certainly continue on the same trajectory.

Stronger force

The notorious laws that opposed genetic technologies, and that made working with modified genes nigh-on impossible, were liberalized by successive governments during the 1990s, and are now in line with those of other Western countries. In 1996, Christian Democrat research minister Jürgen Rüttgers introduced his astonishingly successful BioRegio competition, which brought together academic and industrial scientists to plan the best regional biotechnology initiatives. The time was ripe: the competition evidently succeeded (despite the lack of a tangible prize) because biotechnology has boomed in Germany ever since. The Social Democrats extended the BioRegio idea to an initiative to promote high-tech research in the new Länder.

In 2001, Social Democrat research minister Edelgard Bulmahn introduced laws to make academic salary scales more flexible, so higher salaries could be paid to top researchers, and introduced 'junior professorships', which give young researchers academic freedom at an earlier age. All this will help Germany to compete internationally for top researchers, and will help to break down stifling hierarchies. The main parties all want to extend this flexibility to all levels of research by giving universities the autonomy to hire and fire scientists.

The changes have made Germany an even stronger force in research than it was. Already a respectable second place in European scientific-publication league tables in 1994, behind the United Kingdom, the gap has since narrowed. According to the latest European Union (EU) science indicators, Germany had the largest annual growth in the number of publications of all the large research nations, including Japan and the United States.

The credit for revitalizing Germany's research base cannot be claimed by one political party. Whoever wins the election on 22 September, scientists will notice few major changes in their working lives, although nuclear-energy researchers would doubtless be pleased by the Christian Democrats' commitment to invigorate their discipline after a period in the doldrums under the Social Democrat–Green coalition. The gap in science policy, so wide in 1990, is now minuscule.

Moral issues

But some problematic aspects of Germany's landscape have stubbornly resisted the best efforts of political parties and scientists, to the detriment of German research. In particular, German universities remain unattractive to foreigners, particularly those in eastern Germany, despite their well-equipped labs and attempts to reduce the notorious bureaucratic hurdles for guest scientists. Only around 10% of European postdoctoral researchers holding an EU Marie Curie Fellowship, which enables young scientists to study in another EU country, choose to go to Germany; one-third go to the United Kingdom and 17% to France.

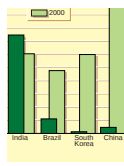
Other obstacles to science are matters of politicians' and voters' principles. Deep moral concerns over issues such as the use of human embryonic stem cells in research will continue to put the brakes on some areas of front-line research, whichever party wins, despite enlightened and sensitive arguments to the contrary (see, for example, *Nature* **412**, 479; 2001). The Christian Democrats' candidate research minister, Annette Schavan, is expected to follow in the footsteps of current minister Edelgard Bulmahn, who has ignored the call of many scientists for a more liberal approach. With all-party support, a law was passed earlier this year banning the use of human embryonic stem cells, with only a few exceptions. And Germany's determined stance has forced a block on the funding of such research in the EU's next Framework research programme.

The main parties are also united in their attitude to research on animals. Although the number of bureaucratic hurdles through which researchers must jump to use animals has been reduced, the Christian Democrats this year abandoned their resistance to a constitutional amendment, which now gives animals rights to be protected by the state. This gives plenty of ground for animal-protection groups to challenge researchers in constitutional courts.

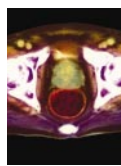
The issue that is most likely to be affected by the election outcome is 'green biotech'. Agricultural biotechnology has always had a rough ride in Germany, but the Christian Democrats and the Social Democrats are both quietly bullish about promoting it. The Social Democrats have been held back by their coalition partner, the Greens. A coalition without the Greens would constitute a government more friendly to genetic modification.

As the political campaigns gather momentum, science is absent from the debates — surprising, perhaps, given the intense concentration of politicians and the media on the issue of human embryonic stem cells throughout the past year. The parties are focusing on the more immediate issues of unemployment, flood damage and Iraq. German scientists, faced with a scientifically neutral choice, can vote according to their political beliefs. ■

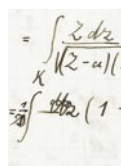




Losing the lead
Indian science slides down publication rankings
p100



On the attack
Europe seeks harmony in fight against cancer
p101



A tidy sum
Physics memorabilia goes under the hammer
p103



Catholic taste
Claims attribute Pope's improved health to papaya
p104

Arrest of AIDS activist underlines China's impending HIV crisis

David Cyranoski, Tokyo

The furtive arrest of a prominent Chinese AIDS activist late last month may have been intended to quell public debate about China's AIDS crisis. But the disappearance of Wan Yanhai — best known for his harsh critique of the government's response to AIDS — has instead cast a spotlight on China's handling of its growing HIV problem.

HIV and AIDS have been spreading at an alarming rate in China. At a press briefing on 6 September, Qi Xiaoqi, director general of the Department of Disease Control at China's health ministry, admitted that the number of people in the country infected with HIV could reach one million by the end of this year — up from 850,000 at the end of 2001.

But many experts think this is an underestimate. A report issued by the Joint United Nations Programme on HIV/AIDS (UNAIDS) in June gave estimates for 2001 of up to 1.5 million and warned that the number of people infected could reach 10 million by 2010.

In much of China, HIV has been spread by sexual contact or through intravenous drug use. But the soaring rates of infection in



Like many in China, 8-year-old Zhang Xiaqing was infected with HIV from a blood transfusion.

central inland provinces such as Anhui and Henan — where some villages have reported HIV infection in up to 80% of adults, according to the UNAIDS report — have largely been caused by the presence of the virus in unregulated blood supplies.

Wan, who was said by his wife and colleagues to be in the custody of security agents two weeks after his unexplained disappearance in Beijing, had been strongly critical of the government's handling of blood collec-

tions. He alleged that local officials were benefiting financially from illegal, unsafe collection of blood through paid donations.

Through his Beijing-based group, the AIDS Action Project, Wan organized petitions to urge the government to provide care for victims. The project's website also attempted to raise awareness of the problem by publishing 'death lists' of those who had died from AIDS in Henan. But in July, the project's office, which operated on a Beijing college campus, was closed down for allegedly not following the rules for non-governmental organizations.

There has been no official announcement that Wan is being detained. But a member of the AIDS Action Project reportedly contacted last week by state security agents says that the agents indicated Wan is being questioned about the contents of an e-mail he sent to a mailing list. The e-mail, which included an official government document describing the number of people suffering from HIV/AIDS in Henan, is classified as a "state secret".

Wan's apparent arrest undermined some recent positive trends in Chinese AIDS policy. In August 2001, for example, the health ministry acknowledged for the first time that

Security worries stifle report on agricultural bioterror

Virginia Gewin

A major report on the threat of agricultural bioterrorism is being delayed as US government officials wrestle with the National Academy of Sciences over whether its release would provide information that could be useful to terrorists.

The academy's report, *Countering Agricultural Bioterrorism*, was commissioned in autumn 2000 and was due to be released in June. But its publication has been bogged down in wrangling between the academy, the US Department of Agriculture and the Office of Homeland Security.

The academy is keen to release the report — which was agreed by a committee chaired

by Harley Moon, an animal-disease specialist at Iowa State University at Ames — in full. But government officials want to cut parts of it.

Members of the committee who had security clearance were provided with at least one piece of classified information during their study, so the study had to undergo a declassification review before it could be released.

Jim Cook, a plant pathologist at Washington State University at Pullman and a member of the committee, who also sits on the academy's governing council, says that the stand-off is unprecedented. "We've never had this kind of difference in perspective — almost a squaring off — between a

government agency and the academy before," he says.

He adds that it reflects a difference in outlook between plant scientists and government officials on countering agricultural bioterrorism. The American Phytopathological Society, for example, has argued that trying to constrain information about plant biology is futile, and that the government should prepare to counter possible attacks. But some government officials favour constraints on information.

As far as the disputed report goes, "the final decision will be the academy's", says Bill Colglazier, its executive officer. He hopes that it will be published later this month. ■

tens of thousands of Henan blood donors were infected through blood transfusions. Last year the government announced a US\$150-million package to establish safe blood-processing and storage centres. And at last week's press briefing, Qi acknowledged China's need for more international assistance in combating the problem. This followed an agreement on 28 June with the United States to collaborate in AIDS prevention and research. Qi also said that China is considering ways to step up the production of medicine for treating AIDS.

"The Ministry of Health is making a lot of effort and has many interesting pilot programmes," says Siri Tellier of the United Nations Population Fund, who headed the group that produced the UNAIDS report. "And they are becoming increasingly open and frank."

But China is still sending out mixed signals. The UNAIDS report accused it of "discrimination, stigma, fear, lack of transparency, and the promotion of information that leads to ignorance and unsafe practices". Xiao Qiang, executive director of New York-based Human Rights in China, brands the treatment of Wan and others as "outrageous".

Such harassment does not necessarily reflect official Chinese policy, some observers say. "The government's programmes are not implemented in a consistent fashion," says Tellier.

♦ www.unaids.org/whatsnew/newaddis/AIDSchina2001update.pdf

Patent office plan to beat its backlog elicits cool response

Kendall Powell, Washington

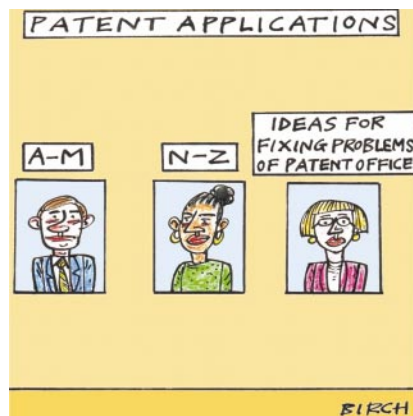
Concern is growing among university and industry officials that proposed reforms at the US Patent and Trademark Office will not solve the office's problems — and may even degrade the quality of the patents it issues.

In response to a request by Congress, James Rogan, the patent office's director, has developed a five-year plan to clear the office's patent backlog. The strategy includes increasing fees, and farming out patent searches — which track down existing patents and other information relevant to an application — to private contractors.

But critics say that the problems could be better addressed if Congress stopped diverting the office's existing fee income to other projects — since 1992, Congress has siphoned off \$672 million in patent fees — and allowed it to hire the examiners it needs.

"I'm encouraged by the patent office's recognition of its problems," says James Stofel, chief technical officer at Eastman Kodak in Rochester, New York. "But I'm frightened by these first steps." He and other industry leaders are concerned at the plan's reliance on untested outside contractors.

Last year, Congress told the office to produce a plan to tackle the existing backlog of 400,000 patent applications and to cut the average of two years it takes to issue a patent.



On 27 August, Jon Dudas, deputy director of the patent office, briefed a National Academy of Sciences meeting on the plan. It includes hiring 3,000 more examiners by 2008, although the office previously estimated that it actually needs 5,500 more. The office also wants to raise its fees by up to 67% next year.

Charles Garriss, a mechanical engineer at George Washington University in Washington DC and a registered patent agent, says that the answer is to hire more examiners. "The rate-determining step of any process is the slowest element," Garriss says. "In this process, it's the very time-consuming step of reading and understanding the applications."

India's scientists agonize over fall in publication rate

K. S. Jayaraman, New Delhi

India has been eclipsed by China as the largest scientific power in the developing world, and is set to fall behind other, far smaller nations, according to an assessment of its scientific productivity.

Subbiah Arunachalam, a science analyst at the M. S. Swaminathan Research Foundation in Chennai, revealed the troubling statistics in the Indian journal *Current Science* (83, 107–108; 2002). He tracked India's publishing record for the period 1980–2000 based on data from the yearly Science Citation Index issued by the Philadelphia-based company ISI.

The total number of papers published by Indian scientists in the journals tracked by ISI for the index fell slightly, from 14,983 in 1980 to 12,127 in 2000, Arunachalam found. But China's output leapt from 924 to 22,061 for the same period, and both South Korea and Brazil expanded their output rapidly (see graph).

In 1973, Arunachalam says, India was the

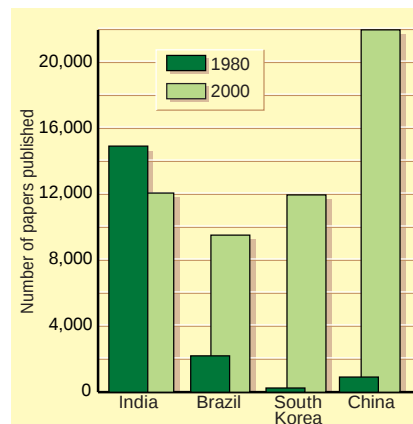
eighth-largest producer of scientific papers in the world, and accounted for nearly half of the developing world's total output of papers. By 2000, it had slipped back to fifteenth in the world ranking.

The analysis has prompted a fresh bout of soul-searching among India's scientific leaders. "Unless we do something serious, I am afraid that we will soon become a third-rate country in science," says C. N. R. Rao, a Bangalore-based chemist and former government adviser who now heads the Third World Academy of Sciences, based at Trieste, Italy.

"All indicators suggest that scientific output is on the decline," agrees P. Balaram, head of biological sciences at the Indian Institute of Science in Bangalore. He attributes the problem to the universities, which he says are suffering from "political interference, diminishing resources, declining faculty quality and deterioration of the academic ambience".

The analysis is particularly depressing

for India as between 1980 and 2000, the number of universities grew from 128 to 231, and research and development expenditure rose substantially.



Falling behind: India has surrendered its scientific dominance in the developing world.

NIH pledges cash for global protein database

Declan Butler, Paris

The United States is turning to European bioinformatics facilities to help it meet its researchers' future needs for databases of protein sequences.

European institutions are set to be the main recipients of a \$15-million, three-year grant from the US National Institutes of Health (NIH), to set up a global database of information on protein sequence and function known as the United Protein Databases, or UniProt.

The NIH is scheduled to announce this week that two-thirds of the grant will go to help maintain two protein databases, Swiss-Prot and TrEMBL. Swiss-Prot is a curated protein-sequence database that strives to provide a high level of annotation including descriptions of function, structure and variance. TrEMBL is a computer-annotated supplement to the main database that contains sequences not yet integrated into Swiss-Prot. The two databases were developed by groups at the European Bioinformatics Institute (EBI) near Cambridge in the UK, and the Swiss Institute of Bioinformatics (SIB), which is based at Geneva and Lausanne.

The remainder of the NIH money will support the Protein Information Resource (PIR) database in the United States, which is kept by the Georgetown University Medical Center in Washington DC, and will now merge with its erstwhile European rivals.

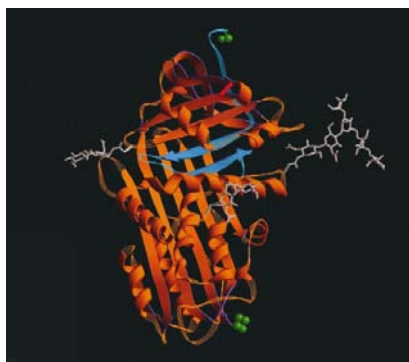
"It's the first time that most of the money from such a large, NIH-funded infrastructure project goes outside the United States," says Rolf Apweiler, Swiss-Prot coordinator at the EBI. "This is remarkable, and shows that Americans recognize that the centre of gravity for protein sequence data is in Europe."

Swiss-Prot began life in 1986 as an augmented version of the PIR database. It has grown to be the world's largest protein sequence database, and, according to some researchers, the most useful — for both the quality of additional editorial and functional information that it carries, and the richness of its links to genomic and other bioinformatics databases.

The NIH is backing Swiss-Prot because it recognizes that the community is best served by focusing efforts globally on one well-developed database, rather than developing a rival, officials close to the decision say.

Under the terms of the grant, PIR will stop maintaining its own database but will assist with the care and feeding of UniProt. Existing data held on PIR will be integrated into TrEMBL and Swiss-Prot. The grant will enable PIR's organization to remain the same size as it is now, but the funding will allow the EBI and the SIB to expand their protein bioinformatics programmes markedly.

For the biologist at the bench, the move is



Sound structure: the planned UniProt database will be a valuable resource for protein research.

likely to help guarantee the availability of a sustainable and reliable source of data on protein sequences and functions, as the

amount of protein data rapidly expands. Database officials say that the new funding will result in a better quality of curated entries, and more sophisticated tools to help users to navigate the large data sets.

But the news that the NIH is backing the project may create some political embarrassment in Europe, given the notorious reluctance of the European Union to support bioinformatics infrastructure (see *Nature* 402, 1; 1999).

"It is ironic that the United States gives us money, whereas the European Union seems to expect such infrastructure to live on manna from heaven," says one European expert in biology and information technology. The US investment inevitably also means that the United States will have a greater say in the control of Europe's bioinformatics infrastructure, the expert points out. ■

EU ponders joint action on cancer

Sally Goodman, Paris

Cancer will kill some 750,000 people in the European Union (EU) this year, making it the area's biggest killer after cardiovascular disease. But the EU has no coordinated research approach, and most of its national programmes are small and fragmented.

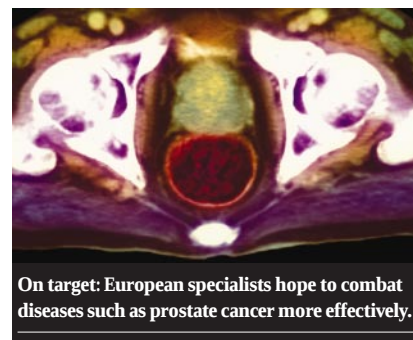
Researchers, government officials and other interested parties will meet in Brussels on 19 September to try to figure out how this can change. Progress is being slowed, many say, by fragmentation, duplication of effort and a lack of harmonization between EU member states' different programmes.

"We can better tackle the complex and multiple causes of cancer by taking advantage of the diversity in Europe," says Elio Riboli, an epidemiologist at the International Agency for Research on Cancer in Lyons, France, who is running a Europe-wide study of nutrition and lifestyle.

Patterns vary hugely across the continent: deaths from breast cancer in Denmark are much higher than in Portugal, for example, yet three times more Portuguese than Danish men die of stomach cancer.

Researchers face obstacles in setting up Europe-wide clinical trials or exchanging tissue samples. Big differences in training create further barriers to cooperation.

Françoise Meunier, director-general of the Brussels-based European Organisation for Research and Treatment of Cancer, says that the number of multinational clinical trials is limited by the effort required to deal with each country's regulatory, ethical and insurance procedures. "Most clinicians cannot free up enough time," she says.



On target: European specialists hope to combat diseases such as prostate cancer more effectively.

Some nations are trying to restructure their own research. Last year, Britain set up a streamlined National Cancer Research Institute, and the country's two largest cancer charities merged this year (see *Nature* 416, 474; 2002). On 9 September, the French government set up a commission to plan its national cancer research programme.

Next week's meeting, organized by the European Commission and the European Parliament, will discuss the establishment of a 'European cancer research area', which would help researchers in non-EU countries to work with their EU counterparts. Research commissioner Philippe Busquin has made cancer one of the priorities of the Commission's 2002–06 Framework programme, which allocates 400 million euros (US\$390 million) to cancer research.

But some researchers will tell the meeting that they want to go further and start a truly European cancer initiative. "Cancer is too important to be left to national governments," says Thomas Tursz, director of the Gustave-Roussy Institute in Paris. ■

Blood banks call for calm over virus scare

Jonathan Knight, San Francisco

As birds and mosquitoes continue to spread West Nile virus (WNV) across the United States, fears have erupted that organ transplants and blood transfusions are also transmitting the virus. But even though WNV can cause a lethal brain inflammation in infected people, screening procedures for blood donors are unlikely to be put in place.

Long endemic to Africa, Asia and southern Europe, WNV arrived in New York in 1999. By summer 2001, it had moved west of the Mississippi. It may now have completed its coast-to-coast march, with a Los Angeles woman returning a positive preliminary test.

WNV is transmitted between birds by mosquitoes, but people and horses can also be infected. The current season is the worst so far — almost 1,000 people have been infected with the virus, and 43 have died.

Questions about the blood supply first

emerged on 31 August, when officials from the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, said they were investigating the case of a Georgia woman given blood from more than 60 donors following a traffic accident. After she died on 1 August, her organs were transplanted to four recipients, all of whom became infected with WNV. One died of encephalitis.

On 5 September, Mississippi health officials said that they were looking into the case of a woman who had received 18 units of blood during surgery several weeks before developing West Nile encephalitis. Investigators are still determining whether any of the blood donors in either case was infected with WNV.

Blood donated in the United States is tested for HIV, the leukaemia viruses HTLV 1 and 2, and hepatitis B and C. In addition, white cells are filtered out to remove herpes viruses. Each test costs up to several dollars per donation,

and new tests are subject to a development and approval process that can take years.

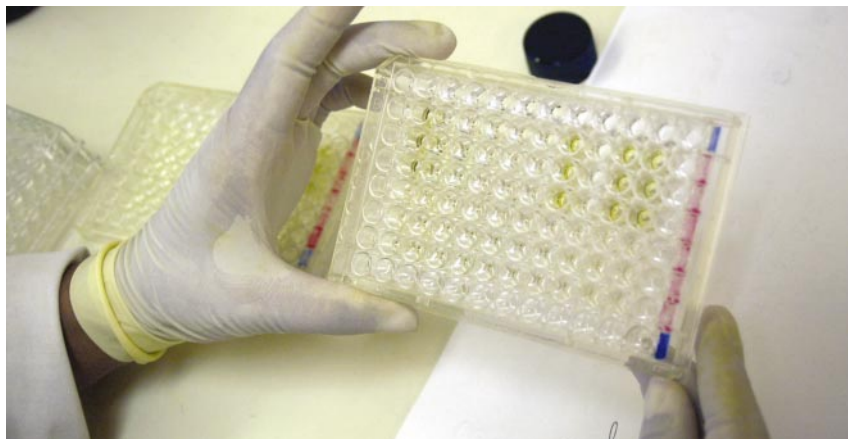
To warrant such a step for WNV, it would have to pose a greater risk than problems already faced by blood banks. From preliminary studies, the CDC estimates that about 1 in 150 infected people develop encephalitis. This suggests that some 150,000 Americans, or 1 in 2,000 people, have been infected this summer, most without knowing it. If a person donated at the peak of infection, the virus might well be transmitted, says Michael Turell of the US Army Medical Research Institute of Infectious Diseases in Fort Detrick, Maryland. But because WNV circulates in the blood for only a few days or weeks, the chances of donors passing it on remain low.

Given these statistics, suggests Michael Busch, vice-president of research at Blood Centers of the Pacific, a San Francisco blood bank, testing for WNV is a lower priority than, for example, detecting low-level bacterial contamination. Bacteria can multiply while blood is stored, and the resulting infections cause about one death for every million units of blood given in the United States.

But the risks may change as WNV spreads. Health officials are on the lookout for the virus, which usually turns up in birds before the first human cases emerge (see *Nature* **417**, 477; 2002). For this reason, California is using some 200 flocks of 'sentinel' chickens to detect the virus; these have so far remained negative.

But on 6 September, Los Angeles health officials reported a positive test for WNV in a woman who had not left the state in some time. Because the test used can crossreact with other viruses, they were awaiting confirmatory tests by the CDC as *Nature* went to press.

♦ www.cdc.gov/ncidod/dvbid/westnile



Keeping track: a technician in Louisiana tests blood samples for signs of the deadly West Nile virus.

Universities urged to get with IT for biology

Kendall Powell, Washington

Biology teachers in universities need to infuse their curricula with more computer science, mathematics and physical sciences, according to leading US scientists.

More cross-disciplinary teaching is needed to keep students up to date with the growing role of computers in biology and the impact of the molecular-genetics revolution, the academics say.

"The way we carry out science is profoundly different from 20 or 30 years ago," says Lubert Stryer, a neurobiologist at Stanford University in California. "But education even at the very best schools has changed very little." Stryer chaired a National Research Council panel that looked into undergraduate biology

education and released its findings in a report on 10 September.

Mounting concerns that undergraduate education in the United States is failing to produce enough young biologists with the skills and motivation to go into research prompted the National Institutes of Health and the Howard Hughes Medical Institute to commission the study.

Although current biology research relies more heavily on computer modelling, information technology and complex instrumentation, undergraduate courses and textbooks have stuck to the facts of life science and not much else, Stryer's panel claims. It also advocates more teaching on the history of science and the process of discovery.

Diane Shakes, who teaches cell biology at

the College of William & Mary in Williamsburg, Virginia, describes the panel's findings as a "noble goal" but sees significant hurdles for smaller teaching universities. She says that hiring interdisciplinary faculty to teach integrated courses would pose problems for small campus departments.

To implement the changes, the study advises using movies, computer presentations, and adding ready-made physics or chemistry modules to existing courses. Team-teaching across academic departments is also recommended.

Stryer hopes that government agencies and private companies will come forward with funds to support teaching workshops and develop educational materials to support the goals of the report.

Looting and vandalism threaten Afghanistan's seed distribution

Natasha McDowell, London

Efforts to revive war-ravaged agriculture in Afghanistan are being threatened by the country's continuing lawlessness, participating scientists say.

"It is difficult to work effectively because some people with political objectives want to create terror and stop others being helped," says Nasrat Wassimi, a plant biologist who is field coordinator of an international attempt to help Afghan farmers (see *Nature* 417, 7; 2002).

Wassimi, who is based in Kabul, admits to fears that recent violence, including an assassination attempt on President Hamid Karzai on 5 September, could be a harbinger of renewed civil war. He says that looting, vandalism and the influence of local warlords are complicating the project's attempts to get seed to farmers for this autumn's planting season.

"You take two steps forward and one back," says Geoffrey Hawtin, director-general of the International Plant Genetic Resources Institute in Rome, part of the Future Harvest network of laboratories that is organizing the effort. "But the alternative is to do nothing."

In spring, the consortium delivered wheat seed to 70,000 farmers in 11 provinces. Now multiplied, this and extra imported seed is being redistributed in time for the larger autumn planting. The project's other priority is the rehabilitation of 22 agricultural stations that have been destroyed over the past 25 years.

Wassimi says that it is essential that traditional and imported varieties of seed are tested so that the most suitable ones can be planted in different regions. Afghanistan's traditional crop varieties are highly diverse, partly because of the many microclimates of its mountain ranges.

After the national seed bank was destroyed in 1992, Wassimi spent two years collecting seed from around the country, and hid it in basements in Jalalabad and Ghazni. But looters destroyed the collection last year, stealing the seeds' plastic containers. At the Darul Aman agricultural station, looters also stole water pumps, and vandals dumped stones into wells, which had to be redug.

Wassimi says he can understand the looting, but not the vandalism. "Whatever destruction you see in Afghanistan, a quarter can be attributed to bombing and shelling, but the rest is vandalism," he observes gloomily. "Some days I feel depressed, but I read a book on positive thinking. That, and wanting to help the farmers, keeps me going."

Early Einstein manuscript set to make a relative fortune

Alison Abbott

Got half a million dollars to spare? If so, you could become the owner of a manuscript that helped to reveal the contribution of Michele Besso, Albert Einstein's lifelong friend and collaborator, to the development of the general theory of relativity.

The 1913 *Einstein-Besso Working Manuscript*, containing more than 50 pages of handwritten text by Einstein and Besso, is part of a private collection of letters, manuscripts and books on twentieth-century theoretical physics that is due to be auctioned by Christie's in New York on 4 October.

Built up by publishing millionaire Harvey Plotnick, the Harvey Plotnick Library on the Development of Quantum Physics and the Theory of Relativity is considered the world's most important private collection of papers on modern physics. It includes around 2,500 items written by more than 100 of the twentieth century's most significant physicists. The total value of the collection is estimated at around US\$2 million.

The *Einstein-Besso Working Manuscript*, which Christie's hopes will fetch between \$500,000 and \$700,000, is the collection's crown jewel. Besso, a Swiss engineer, was Einstein's most dependable sounding-board. The manuscript includes mathematical calculations, carried out by Besso under Einstein's guidance, which test whether early versions of Einstein's general-relativity equations could account for an anomaly in Mercury's orbit that Newton's theory of gravity couldn't fully explain.

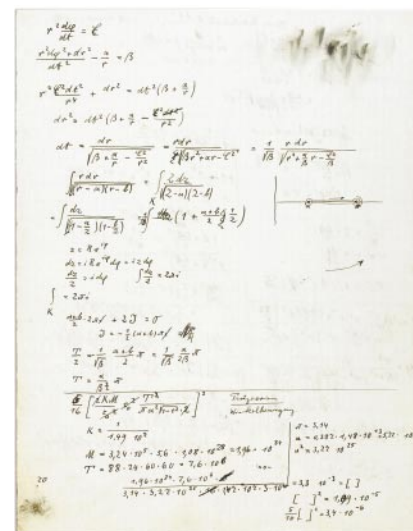
The equations failed to solve the problem, but reformulated equations, finalized in 1915, were eventually successful. Along with other items of correspondence still held by the Besso family, the manuscript has helped science historians to piece together the working relationship between Einstein and Besso.

"Besso contributed more to the intellectual development of Einstein's general theory of relativity than has previously been realized," says Jürgen Renn, a director of the Max Planck Institute for the History of Science in Berlin, and co-editor of the fourth volume of *The Collected Papers of Albert Einstein* (Princeton University Press, 1995), which includes a reproduction of the *Einstein-Besso Working Manuscript*. Renn, together with Michael Janssen, a science historian at the University of Minnesota, is preparing an article based on another previously unpublished manuscript by Besso, which underlines the significance of Besso's work in the development of the equations.

Other items in the Plotnick library include the doctoral thesis of Henri Becquerel, who



Michele Besso (above right, with his wife Anna) worked with Einstein on his relativity paper.



discovered naturally occurring radioactivity, and that of Marie Curie, discoverer of the radioactive elements polonium and radium. There is also an offprint of Wilhelm Röntgen's 1895 announcement of his discovery of X-rays, and three handwritten letters by Werner Heisenberg to the Dutch physicist Samuel Goudsmit, discussing Goudsmit's discovery of electron spin.

Plotnick, who lives in Chicago, started his collection around 20 years ago. "I've always been interested in the history of physics," he says. Having learnt enough about the development of quantum physics, Plotnick is selling in order to concentrate on his new interest in early Islamic art.

Most historians prefer important documents to be held by public institutions to safeguard access, but they accept that private ownership doesn't necessarily hinder their work. Plotnick has already given many historians access to the documents.

ALP/EMILIO SEGRE VISUAL ARCHIVES

CHRISTIE'S NEW YORK

Zimbabwe reaches deal with UN on transgenic food

Harare The Zimbabwean government has dropped its opposition to accepting genetically modified food as aid, according to officials at the United Nations' World Food Programme (WFP).

About six million Zimbabweans face food shortages caused by drought and the government's policy of seizing farms owned by whites. Some Zimbabwean politicians had wanted to reject offers of transgenic grain from the United States. They feared that some seeds would be planted, potentially harming future exports to Europe, where consumers are reluctant to buy food containing transgenic ingredients (see *Nature* **418**, 571–572; 2002).

But President Robert Mugabe said last week that his country will now accept 17,500 tonnes of yellow maize donated by the US government. The grain will be milled so that the seeds cannot be planted.

"We are very encouraged by this and hope the agreement will be formalized shortly," says Luis Clemens of the WFP. Officials in Zambia, which is also facing food shortages, have yet to accept transgenic food, although the programme hopes that Zimbabwe's decision will encourage them to do so.

US gives clearance for first private Moon landing

Washington The first privately funded mission to the Moon has been given clearance for take-off by the US government.

TransOrbital, a space-exploration company based in La Jolla, California, announced on 22 August that it has won permission for its TrailBlazer probe to map the Moon's surface, photograph the Earth and deposit a time capsule on the Moon. As a company based in the United States, it had to obtain its permit from the US government. The mission, scheduled for a June 2003 launch from the

Baikonur Cosmodrome in Kazakhstan, will be funded by corporate sponsorship and by selling video footage and images.

Obtaining the permit took more than two years — the company had to prove it would not contaminate the Moon with biological material, pollute the surface or disturb any historical landing sites.

♦ www.transorbital.net

Anger at plans to cut French research funds

Paris President Jacques Chirac is facing a storm of criticism following revelations that France's research budget is to be cut by 1.3% in 2003. According to official figures leaked to *Le Monde* newspaper last week, the cuts will include the loss of 50 jobs from publicly funded research.

The figures, contained in a letter from Prime Minister Jean-Pierre Raffarin to research minister Claudie Haigneré, is expected to be confirmed at a meeting of ministers on 25 September.

The news has annoyed researchers' trade unions, who are accusing the recently re-elected Chirac of reneging on his election promise to ensure that 3% of gross domestic product was invested in research by 2010. It had been estimated that public spending on research would need to increase by 4.2% annually if this target was to be reached.

The unions are also angry that the new right-wing government is apparently not going to follow the previous socialist government's plan to create 1,000 new research posts before 2004. Critics say the decision echoes the initial years of Chirac's first term as president with a right-wing government, in 1995–97, when the research budget and science jobs were cut.

Fruitful meeting between the Pope and Montagnier

Paris Luc Montagnier, head of the laboratory at the Pasteur Institute in Paris that discovered the AIDS virus in 1983, has made an unconventional attempt to treat Pope John Paul II, who is thought to be suffering from Parkinson's disease.

Montagnier met the Pope in June to discuss the AIDS pandemic in Africa, and to lobby for a shift in the Catholic church's opposition to condom use. But he also offered the Pope an experimental concoction of fermented Asian papayas, which he claims can stimulate the immune system and also mop up free radicals. Montagnier believes that free radicals play a role in many chronic diseases and could be used to treat conditions such as Parkinson's. The benefits of papaya have not yet not been proven, however.

Observers have remarked on a spectacular improvement in the Pope's health, especially his speech, during his recent trips to Canada



Step on the juice: could papaya be the source of the Pope's recently improved health?

and Poland. But the Vatican has never confirmed that the Pope has Parkinson's, nor that he took the papaya prescription.

Two fired in wake of anthrax investigation

Washington Former biodefence researcher Steven Hatfill, who was questioned by the FBI over last autumn's anthrax attacks, has been fired by Louisiana State University (LSU). Stephen Guillot, director of the National Center for Biomedical Research and Training at LSU where Hatfill worked, has also been removed from his job.

Hatfill, who has been described by government officials as a "person of interest" in connection to the anthrax attacks (see *Nature* **418**, 808; 2002), was fired on 3 September. He was subject to intense media and government scrutiny during the FBI's investigation, but no solid link has emerged between him and the attacks.

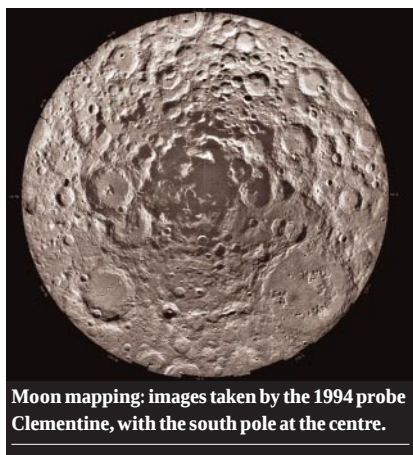
Newspaper reports claim that the Justice Department, one of the LSU centre's primary funders, demanded that Guillot remove Hatfill from all the projects that it funds, but the centre refused to confirm why Hatfill had been fired. Administrators at the university were equally tight-lipped over the firing of Guillot from his position a day later.

Star spotting leads to claim that planet is just an illusion

Washington At least one of the 100 or so planets discovered around other stars is an impostor, according to new research.

Astronomers claimed in 1999 that shifts in the frequency of light from the star HD 192263 were due to the pull of a large planet. But a paper to be published in the 1 October issue of *Astrophysical Journal Letters* finds that the light varies on a regular cycle. Gregory Henry of Tennessee State University, Nashville, and his team attribute this to the rotation of large 'star-spots', cooler regions associated with magnetic activity, on the star's surface.

The same phenomenon may eventually disqualify other purported planets, they add, but it only a few as it just occurs on stars with very active surfaces.



Moon mapping: images taken by the 1994 probe Clementine, with the south pole at the centre.

An out of body experience

Reproductive and neonatal medicine have cut the minimum time that fetuses need to spend in the womb. Are completely artificial pregnancies in the offing? Jonathan Knight investigates.

Scott Gelfand wants us to learn a lesson from cloning. When Dolly the sheep burst onto the public stage in 1997, ethicists and politicians were caught by surprise. Five years later, legislation around the world is still being updated. Gelfand, director of the Ethics Center at Oklahoma State University in Tulsa, is determined not to let the same thing happen with the artificial womb. So in February, he hosted a conference entitled: 'The end of natural motherhood? The artificial womb and designer babies.'

No one in attendance believed that Aldous Huxley's *Brave New World*, in which fetuses were nurtured in the laboratory rather than the womb, is about to become reality. But research into how to help premature babies has already allowed those born as young as 24 weeks into pregnancy to live. *In vitro* fertilization (IVF) now allows embryos

to be cultured for up to three days before being implanted, and researchers believe that longer times may be possible. If current methods were pushed to their absolute limit, normal human gestation could, in theory, be halved to 20 weeks, although not without probable harm to the fetus. "At some point the two ends will connect," Gelfand claims. "We need to start thinking about this now."

The researchers behind these projects say that they are not trying to turn pregnancy into a laboratory procedure and, in any case, the technology to nurture a fetus's normal development throughout gestation remains in the realms of science fiction. But even at its present stage, the research is posing some tough ethical questions that impinge on the contentious ground of abortion politics.

Premature problems

So far, most attention has focused on helping premature infants to survive. The lungs of babies born more than four weeks premature usually lack a foamy surfactant liquid that stops the organs from collapsing between breaths. This can often be overcome using surfactant chemicals, together with mechanical breathing aids, which can help infants born as young as 24 weeks to survive.

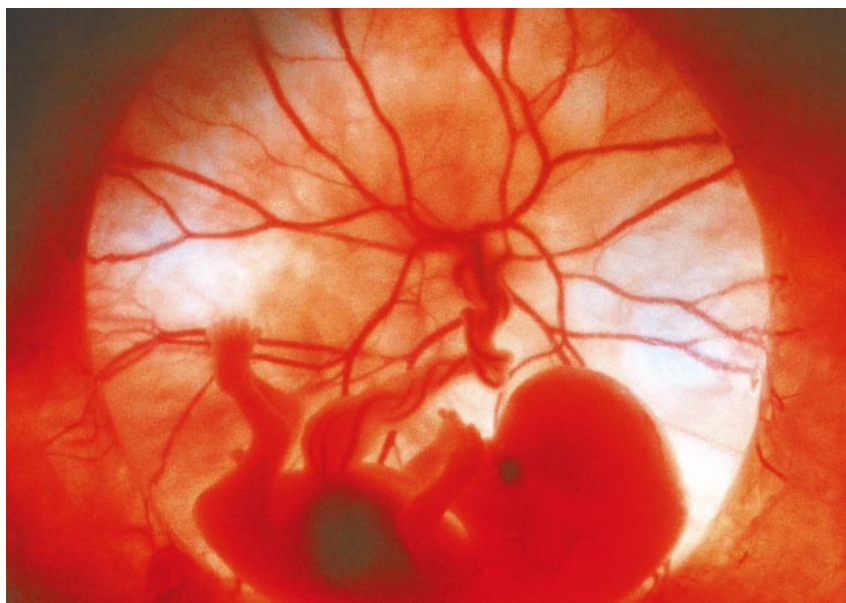
But very premature babies can still end up mentally impaired, and mechanical ventilation can damage the lungs. Thomas Shaffer, a neonatal physiologist at Temple University in Philadelphia, believes that liquid ventilation could help. In the womb, the fetus receives oxygen through the umbilical cord, and its lungs are filled with protective amniotic fluid. But at about 23 weeks, a fetus's lungs are ready to take in oxygen. So, in theory, premature babies could get oxygen from a fluid in their lungs. This should be a gentler alternative to forced ventilation and might deliver oxygen to the brain more efficiently than other

methods, which in turn might reduce the chance of brain damage occurring.

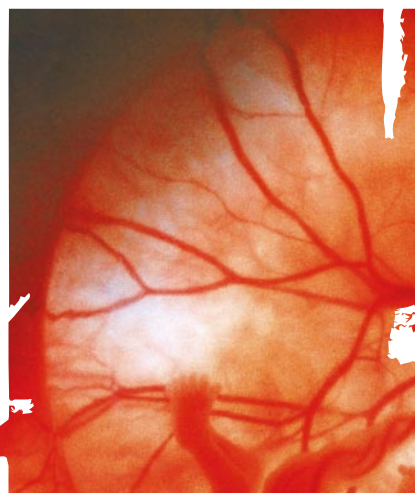
Shaffer began testing liquid ventilation in the early 1980s. He filled the lungs of 18-week-old lamb fetuses with a perfluorinated hydrocarbon that had been infused with oxygen. These inert organic compounds have a remarkable ability to carry oxygen, and are also being experimented with as substitutes for blood transfusions. Shaffer found that the fluid provided a better flow of oxygen to the blood than forced ventilation¹.

Then, by accident, his supplier sent a ewe that was only 14 weeks pregnant, which Shaffer only discovered when he was removing the fetus. To his surprise, the animal's immature lungs responded just as well as those in the older lambs², suggesting that the technique could be used in very premature infants.

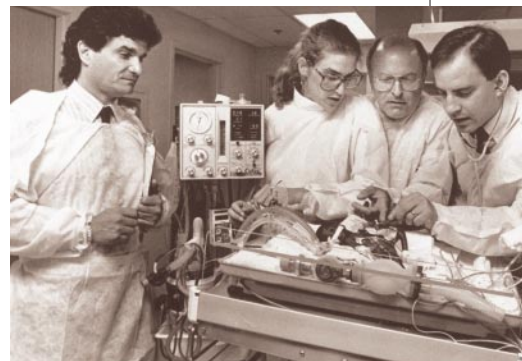
In 1989, Shaffer attempted a short application of liquid ventilation on three human babies at St Christopher's Hospital for Children in Philadelphia, the most premature of which was just 23 weeks old. The infants, who had failed to respond to other treatments, survived for less than a day after the



EDELMANN/SPL



Brave new world? Scott Gelfand says we must debate the ethics of artificial wombs now.



Thomas Shaffer (far left) and his team hope that fluids could help premature babies to breathe.

TEMPLE UNIV. SCHOOL OF MEDICINE

fluid ventilation was removed, but two received more oxygen to the blood while they were given liquid instead of air, and all three had improved lung expansion³.

A 1996 clinical trial also provided positive results. Thirteen infants, born after 24–34 weeks with severe breathing difficulties, were given oxygenated liquid for between four hours and three days. Although none had been expected to survive, seven were discharged from hospital and appeared to be healthy and normal months later⁴. Those that Shaffer has followed up since are still doing fine. But he cannot find a drug company willing to continue with the clinical experiments. Shaffer suspects that the small size of the neonatal market deters firms from taking on the costs of the clinical trials.

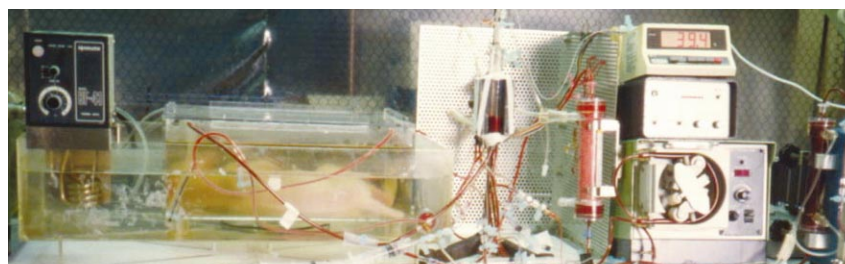
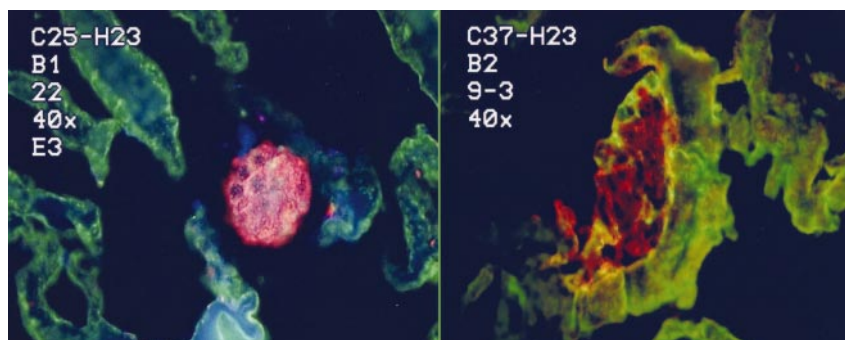
Direct line

Other researchers have worked on systems that could help babies born even earlier by oxygenating the blood directly. Several groups pursued this approach in animal experiments in the 1960s and 1970s, before assisted ventilation became established. But Yoshinori Kuwabara, then at the University of Tokyo, continued in the hope that the method might treat very premature infants and could aid research into fetal physiology.

In 1990, his team used what was essentially an artificial placenta to keep mid- to late-stage goat fetuses alive. Catheters carrying nutrient-enriched blood were threaded into the umbilical arteries and veins of the goat fetuses, which were held in a tank of artificial amniotic fluid. Early attempts were complicated because the fetuses tended to pinch or pull out the catheters as they twisted and kicked, so Kuwabara paralysed them with a muscle relaxant. This allowed two goats extracted from the womb three weeks early to survive until their normal term⁵. Unfortunately, the relaxant prevented the fetuses from developing muscle tone, and they could not stand or breathe unassisted. When the researchers removed the ventilator after four weeks, the goats died within hours.

The project moved to Juntendo University in Tokyo in 1992, and continued there after Kuwabara's death in 2000 — although the researchers are now using the technique to study the flow of blood around the fetal brain. Plans to develop a treatment for premature babies are on hold until they work out how to solve the catheter problem. But Nobuya Unno, a member of Kuwabara's team who is now at Nagano Children's Hospital, argues that this type of incubation could potentially be used to help babies born at about 20 weeks to survive until their lungs are mature enough to allow fluid ventilation to be used.

Establishing an artificial placenta for babies born earlier than that is likely to be difficult. Before 20 weeks, umbilical blood vessels are too narrow to accept catheters. The blood must also be infused with an anticoagulant to



Culturing IVF embryos with cells from the uterus (top left) mimics the womb, and the embryo embeds in the cells after six days (top right). Meanwhile, premature goat fetuses can be kept alive using an artificial placenta (bottom).



stop it clotting in the machinery, and anticoagulants tend to cause brain haemorrhaging in very immature fetuses. Although the barriers are not insurmountable, Unno says that he has no interest in pursuing such an approach.

Fertile ground

More recently, IVF research has been chipping away at the other end of gestation. Hung-Ching Liu, an embryologist at Cornell University's Center for Reproductive Medicine and Infertility in New York, is studying how the embryo attaches itself to the lining of the womb, known as the endometrium. She uses an IVF technique in which a fertilized human egg is cultured with endometrial cells before implantation, but she has modified the process so that the endometrial cells are first grown in a collagen matrix contained in a plastic well less than a centimetre wide. This causes the cells to form multilayered plugs similar to the tissue of the uterine lining.

In unpublished work, Liu found that the fertilized eggs attached themselves to the plugs of endometrial cells at six days post-fertilization — just as they do in the womb. Liu hopes that growing embryos in such uterus-like conditions before implanting them will improve the success rate of IVF. She plans to extend her experiments to 14 days after fertilization to study how the placenta begins to grow, but says she has no interest in going beyond this stage.

As with the Japanese work, Liu's experi-

ments have generated media claims that an artificial womb is on the horizon. All of the researchers dismiss this as speculation, but for Gelfand, it flags up the need to begin talking about the ethical issues. Delegates at his conference asked whether the technology could, for example, offer infertile couples who do not want to use a surrogate mother a chance of having their own baby. But under what circumstances should they be allowed to do so?

Attempts to boost the survival of premature babies are already raising difficult questions about time limits on abortions. In Britain, for example, a woman can under most circumstances legally have an abortion no later than 24 weeks into her pregnancy. This limit was reduced from 28 weeks in 1990, in part because medical advances meant that infants born at this stage could survive. Should further advances rescue even younger fetuses, the pressure to reduce the age again could mount.

Given the incendiary nature of the abortion debate, research at the frontiers of neonatal medicine seems certain to keep Gelfand and other ethicists busy — even if the spectre of Huxley's novel remains a distant prospect. ■

Jonathan Knight writes for *Nature* from San Francisco.

1. Shaffer, T. H., Douglas, P. R., Lowe, C. A. & Bhutani, V. K. *Pediatr. Res.* **17**, 303–306 (1983).
2. Shaffer, T. H., Tran, N., Bhutani, V. K. & Sivieri, E. M. *Pediatr. Res.* **17**, 680–684 (1983).
3. Greenspan, J. S., Wolfson, M. R., Rubenstein, S. D. & Shaffer, T. H. *J. Pediatr.* **117**, 106–111 (1990).
4. Leach, C. L. et al. *N. Engl. J. Med.* **335**, 761–767 (1996).
5. Unno, N. et al. *Artif. Organs* **17**, 996–1003 (1993).

A window of opportunity

Enthusiasts for European scientific integration believe the time is ripe to launch a new independent research agency. Quirin Schiermeier examines the case for a European Research Council.



"There is nothing more powerful than an idea whose time has come."

With this quote from the nineteenth-century poet Victor Hugo, leader of the French romantic movement, the Strasbourg-based organization Euroscience drafted a position paper in July backing the creation of a European Research Council (ERC). Euroscience's website is now soliciting views on the topic.

Euroscience is an alliance of more than 1,200 scientists, formed in 1997 to promote research at the European level. No surprise, then, that it endorses plans to create an independent pan-European body to support research across the disciplinary spectrum, driven by scientific excellence rather than political pressures. But backing for an ERC extends beyond the ranks of the scientific community to include senior figures within national funding agencies and politicians from the member states of the European Union (EU).

So what, exactly, is the idea? And has its time really come, or does the choice of a Hugo quote as a rallying call betray an incurable romanticism among the ERC's supporters? After all, some people have been calling for the creation of a European equivalent of the US National Science Foundation (NSF) for years, with little to show for their efforts.

The answers to these questions should move into sharper focus next month, when scientists and research managers from across

the continent gather in Copenhagen to discuss the idea of creating an ERC, at a meeting hosted by Denmark's research councils.

Research leaders argue that European nations need to pull together more effectively if they are to compete with the research powerhouse that is the United States. Indeed, fears about European competitiveness lie at the heart of a proposal by the EU commissioner for research, Philippe Busquin, to coordinate national science policies under the umbrella of a European Research Area (see *Nature* **413**, 768–770; 2001). Many observers, however, believe that an independent ERC is needed to make any real headway in integrating Europe's fragmented efforts in basic research.

Gathering speed

The ERC idea started gaining its current momentum in Scandinavia. At an informal meeting on European science policy held in Stockholm in April, Pär Omeling, director-general of the Swedish Research Council, outlined a model for a multidisciplinary council financed by the EU but acting autonomously. It would concentrate on fundamental and strategic research, moving swiftly and basing funding decisions on scientific excellence judged by peer review.

ERC supporters contrast this vision with the EU's Framework research programme, which is primarily concerned with making Europe more economically competitive and inevitably gets influenced by the political

agendas of member states. The Framework programme is widely criticized for responding sluggishly to new scientific opportunities — not surprisingly, as its budget is set in five-year chunks in a process that requires agreement among the EU's member states and approval by the European Parliament.

Danish officials are also supportive of the ERC idea. Though Scandinavian countries are traditionally sceptical about the value of European-level institutions, they are enthusiastic about the ERC — not least because their researchers are likely to compete successfully for funding if quality is the driving factor. With Denmark now occupying the EU presidency, which rotates every six months, the ERC is in the political spotlight.

At the same time, enthusiasts spy an opportunity in the deliberations of the European Convention, a high-level panel chaired by former French president Valéry Giscard d'Estaing. The convention is preparing a far-reaching reform of EU administrative structures by 2004, when 11 more nations are due to join its existing 15 members. As it reconsiders the EU's current activities, including the huge sums devoted to agricultural subsidies, it may be possible to throw the ERC proposal into the mix. "There is a window of opportunity," suggests Portugal's former research minister José Mariano Gago, head of the Portuguese Laboratory for Particle Physics in Lisbon and a long-time supporter of greater European integration in science.



José Mariano Gago: launch a voluntary pilot scheme.

Among ERC enthusiasts, however, there is a diversity of views on the best starting point, in terms of disciplinary scope, modes of funding, and governing statutes. Last, but by no means least, is the key question of where the money should come from. The wide array of competing ideas should provide ample scope for discussion at the Copenhagen meeting, which opens on 7 October.

Perhaps the most provocative proposal comes from Hans Wigzell, director of the Karolinska Institute in Stockholm, Sweden's premier centre for biomedical research. He advocates splitting the Framework budget — 17.5 billion euros (US\$17.2 billion) for the current five-year programme — straight down the middle, allocating half to the ERC. By contrast, Ernst-Ludwig Winnacker, president of the DFG, Germany's main granting body for basic research, suggests that the money should come from national agencies, which could initially contribute 0.5% of their budgets — giving a nascent ERC, involving perhaps half-a-dozen agencies, some 25 million euros per year. Other experts suggest various mixtures of EU and national funding, perhaps supplemented by donations from industry.

Some ERC proponents want it to operate like the NSF, supporting a range of activities from fellowships through project grants to funding for large facilities. Others suggest a series of ERCs in individual disciplines, possibly affiliated to existing bodies such as the European Molecular Biology Laboratory in Heidelberg or CERN, the European laboratory for particle physics near Geneva.



Ernst-Ludwig Winnacker: wants national funding.

Richard Sykes, rector of London's Imperial College, who leads a high-level working group of the European Science Foundation (ESF) considering the ERC idea, suggests that it should initially fund the cream of Europe's postdoctoral researchers to work abroad for four or five years. Making 1,000 awards per year would require an annual budget of some 200 million euros, he says.

For many working scientists, an independent ERC is an appealing idea. Some see an opportunity to overcome problems at the national level in access to major research facilities. Katherine Richardson Christensen, an oceanographer and vice-rector of the University of Aarhus, says Danish agencies

The Polarstern: star vessel for Antarctic research, but few have access to it.



cannot obtain access to Europe's only ship capable of conducting multidisciplinary research in Antarctic waters: Germany's *Polarstern*. "Europe has a whole fleet of small research vessels, but only a few scientists get access to European resources," she says.

In candidate EU members such as Poland, where research funding is scarce, the ERC is seen as a lifeline for scientists who can compete with Europe's best. "This would be most welcome in Poland, where not even a Nobel candidate has a chance of getting serious money," says Jerzy Langer, a solid-state physicist at the Polish Academy of Science's Institute of Physics in Warsaw, and vice-president of Euroscience. He suggests that the ERC should support fellowships to allow talented researchers of all ages to stay in their home countries.

Doubts and downsides

But Ian Halliday, head of Britain's Particle Physics and Astronomy Research Council, questions whether an ERC will be able to



Jerzy Langer: a lifeline to countries such as Poland.

meet all the hopes being pinned on it. "Small countries may look favourably on the ERC," he says, "but do they really know how hard it would be to compete for funds?"

Halliday backs the goal of a European funding agency supporting scientific excellence. But he warns that care will be needed to ensure that political pressures don't distort that mission, and that spending is concentrated on initiatives that require support at the European level. "Taking money away from national councils, just to spread it all over Europe, would create no added value," says Halliday. "It is easy to say that the ERC would fund only the very best, and automatically do better than the Framework programmes. But who guarantees that it does not get political?"

The ERC's supporters argue that it should be a self-governing body, run by scientists for scientists. Meanwhile, the Strasbourg-based ESF, an association of national research agencies which already runs pan-European

research projects funded by its members, suggests that it could play a role in establishing and running a fledgling ERC.

But any body using money from the EU's central coffers and from national research agencies would have to be accountable to the EU and participating states — which raises tough questions about the extent to which its paymasters would influence its priorities.

"There are no simple answers," Gago concedes. There is also the obstacle of convincing politicians and national agencies, particularly in the major scientific nations, to sanction any transfer of funds. Though Winnacker is an ERC enthusiast, for instance, the idea does not yet have the DFG's formal backing.

Nevertheless, Gago remains convinced that the political climate is more favourable than ever before. "I believe all EU member countries are open to the matter," he says. Though few observers expect the Copenhagen meeting to produce a single clear vision for the ERC, the idea's supporters hope it will



Ian Halliday: who guarantees it will not get political?

at least narrow the range of possibilities, and show a willingness by scientists and research administrators to get involved.

To seize the current window of opportunity, seasoned observers of the European scene say the ERC's proponents will have to launch at least a pilot scheme — such as Sykes's proposed fellowships — within a year.

"The most simple way would be for a federation of national research agencies to launch, on a voluntary basis, one or more common programmes with their own budgets," says Gago. ■

Quirin Schiermeier is *Nature's* German correspondent.

Euroscience's ERC debate

► www.euroscience.org/WGROUPS/SCIENCE_POL/erc.htm

Copenhagen meeting

► www.forsk.dk/dkeuformand/information.pdf

European Convention

► european-convention.eu.int

European Science Foundation

► www.esf.org

Can commercial protection be good for research?

It's unscientific to let market forces dictate a change in the way science is carried out.

Sir — Ari Patrinos and Dan Drell (*Nature* **417**, 589–590; 2002) recommend that journals embrace the commercial realities facing science, suggesting that arrangements such as *Science* allowing some data restriction are necessary for progress. Although they accept that open access to research data has been crucial to scientific success, their core argument is that too much data will remain locked in the private sector unless journals become more accommodating. Patrinos and Drell's suggestion is that publicly funded information should be unrestrictedly available, whereas privately funded findings can be legitimately restricted for an agreed time.

Such reasoning may be the start of a slippery slope leading to different standards and treatment for privately funded, profit-making science. The growing trend towards public–private partnerships will grease the slope and stud it with inconsistencies. Will the public's share of data be embargoed with the

private? Will the length of the embargo be determined by the ratio of public and private money? Will the time of data release be negotiated by authors with publishers on an individual basis? After all, offsetting the credit reward from early disclosure against the loss of privileged access to one's own data is as important to the career of a publicly funded scientist as to the profit of a corporation.

Even more general than these considerations is the role of data disclosure in scientific success. The tension between openness and protection of novel findings is a key feature in the dynamics of collective scientific practice. The way scientists have historically balanced competition and cooperation has led to generations of scientific achievement. More data may indeed be gained for all from special publication arrangements, but will this be sufficient compensation for the loss of a mechanism that has been crucial in the extraordinary accomplishments of many disciplines?

Both supporters of early, full data release and advocates of partial commercial protection argue that their strategy leads to more scientific progress. We agree with Patrinos and Drell that norms of 'free access' and 'total disclosure' are unrealistic: indeed, pre-publication secrecy has always been part of successful science. Nevertheless, accepting a trend of increasing market orientation as a policy guide is unscientific. How science works is itself a scientific question, not an ethical or pragmatic one. Surely, scientific inquiry should rule on the advisability of overthrowing conventions associated so strongly with scientific success. It is remarkable that the effects of radical changes in institutional settings and reward structures for scientists have not been more extensively researched.

Maureen. A. O'Malley, Andrew J. Roger, W. Ford Doolittle

Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Industry scientists look for benefits, not risks

Sir — The Correspondence "Conflicts around a study of Mexican crops" (*Nature* **417**, 897–898; 2002) was an excellent reminder to readers of what it means to be a scientist in a value-laden world. But surely nobody can be surprised that those involved in this controversy are accusing each other of conflicts of interest, or that indeed they have such conflicts?

I would not expect scientists in the genetic-modification industry to publish experiments showing whether genetically modified organisms pose a risk to the environment or human health, because their priorities are improvements and applications. The search for risk is therefore left to researchers who have some reason, whether ideological or any other, to hypothesize that there might be a risk.

What seemed to me the most important aspect of the Correspondence debate was not the possible conflicts of interest themselves, but the honesty of all the participants. Conflicts of interest by authors of publications can be handled by openly declaring them, yet the potential honesty of the authors is less easy to decipher. This is what we should worry about, because in order to survive as a profession we need to ensure that, in the words of your Opinion article

(*Nature* **418**, 111; 2002): "trust in science remains deservedly high".

Maria J. Hötzel

LETA – Laboratory for Applied Ethology, Departamento de Zootecnia e Desenvolvimento Rural, Universidade Federal de Santa Catarina, Rodovia Admar Gonzaga, 1346, Itacorubi, Florianópolis, SC 88.034-001, Brazil

Public-access group supports PubMed Central

Sir — Your News story "Public-access group plans journals" (*Nature* **418**, 805; 2002) reported one of us (M.B.E.) as saying on behalf of the Public Library of Science (PLS) that PubMed Central is "woefully inadequate" in meeting researchers' needs. In fact, the PLS strongly supports PubMed Central (<http://pubmedcentral.nih.gov/>) and its laudable efforts to create a digital archive of the scientific literature that is freely accessible and fully searchable. What was criticized as "woefully inadequate" is not PubMed Central itself, but publishers' participation in it: since its founding in 1999, relatively few publishers have taken advantage of this opportunity to serve the scientific community better.

PLS (www.pubmedlibraryofscience.org) was formed nearly two years ago to show publishers that the scientific community supports PubMed Central and other free,

full-text public libraries of scientific literature. Although it has received strong, broad-based support from scientists and the public, most publishers — including those of *Nature* — are, in our opinion, resisting the will of the scientific community and are the major obstacle to improving access to the literature. It is this failure, not any dissatisfaction with PubMed Central itself, that leads us to conclude that the scientific community will have to take the initiative to create new journals that enable immediate and unencumbered access to published reports through public resources.

It is not the case, as your News story states, that PLS withdrew its boycott. It continues to urge scientists to show their support for the principle of open access to the scientific literature by publishing exclusively in journals that make their content freely available through archives such as PubMed Central.

Michael B. Eisen*, Patrick O. Brown†, Harold E. Varmus‡

**Life Sciences Division, Lawrence Berkeley National Lab and Department of Molecular and Cell Biology, University of California,*

1 Cyclotron Road, Berkeley, California 94720, USA

†Department of Biochemistry and Howard Hughes Medical Institutes,

Stanford University School of Medicine,

300 Pasteur Drive, Stanford, California 94305, USA

‡Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

Biology on the global scale

All of the Earth's systems — including life — are intimately connected.

The Earth's Biosphere: Evolution, Dynamics, and Change

by Vaclav Smil

MIT Press: 2002. 360 pp. \$32.95, £21.95

Peter Westbroek

No object in the Solar System is more enigmatic than our own familiar planet. At least four of its attributes stand out as unique, as far as we know. The slow mill of tectonics pumps the rocks around from the deep Earth to the surface and back; two-thirds of the planet is covered with liquid water; the atmospheric composition is far removed from thermodynamic equilibrium; and, most stunning of all, there is abundant life — a biosphere. We feel that these remarkable attributes are intimately connected, but are poorly informed about these relations and by what principles the system evolved. Earth dynamics, in particular the role of life, is hot stuff at present, Earth system science and geobiology being examples of new disciplines specifically concerned with this problem. So Vaclav Smil's overview of what we know about the biosphere is timely.

To those reading *Nature* today, it may come as a surprise that only a few decades ago Earth dynamics was overwhelmingly considered as resulting from purely physico-chemical interactions. Even Edward Suess, who coined the term 'biosphere' in 1875, thought of life as basically independent of the rest of the planet. In retrospect, it seems as though the geologists first had to settle the more basic aspects of their subject before they could concentrate on the intricacies of geobiology.

This may explain why the Russian scientist Vladimir I. Vernadsky, who had already set the stage for modern geobiology in the first half of the twentieth century, could be ignored for so long in the Anglo-Saxon part of the world. It is one of the great merits of Smil's book that it puts this record straight. The volume opens with a vivid description of Vernadsky's eventful life and career, stressing the enormous achievements of this great scientist. *The Earth's Biosphere* may be read as an update of Vernadsky's key book *The Biosphere*, which only appeared in English as recently as 1998 (Copernicus Books, translated by D.B. Langmuir).

The subject is immense, encompassing the present state of life and its environment, as well as the history of the system over almost 4 billion years. To deal with a field of this complexity, two options are open. One can either try to organize the different aspects around a single vantage point, or just give a kaleidoscopic description of the state



Tree of life: tropical forests support biodiversity by providing habitats for a huge range of insect species.

of our knowledge. The first approach has the advantage of coherence and perspective. Things fall into place and there is an argument that one can agree with or not. James Lovelock's book on Gaia (*The Ages of Gaia: A Biography of Our Living Earth*, Norton, 1988) is an example. It is easily accessible even to a lay audience and, although the message has been criticized, the book has sparked a vivid interest in a systems approach to the Earth and to global geobiological feedbacks.

Another example is S. A. L. M. Kooijman's *Dynamic Energy and Mass Budgets in Biological Systems* (Cambridge University Press, 2000; 2nd edition), which presents a theory of biological organization at the level of the individual organism. Based on a very limited set of assumptions, this theory has a high predictive value and covers an amazing variety of natural phenomena. The ambition is to extend the organismal models to the molecular, ecological and eventually global levels of organization.

Smil's book is closer to the descriptive end of the scale. It covers almost the full range of relevant subjects, while many of the actual relationships between them remain open-ended. Discussions of the physical aspects of the system include the generation and output of energy by the Sun, plate tectonics, oceanic and atmospheric current systems and global climate. Biological subjects include molecular processes, organis-

mal biology, ecology and global phenomena. Thus there are passages on basic molecular-biological concepts and the origin of life, metabolic pathways, microbial ecology, and biogeochemical cycling. The global dimension includes discussions on the deep hot biosphere and biospheric mass and productivity. The book ends with the impact of civilization on the biosphere.

The above is only a selection, and the whole text becomes a bit of a pot-pourri. I feel that this not only reflects the author's jumpy mind, but also the state of the art. Nevertheless, the book offers updated reviews of many of these subjects, and most of its statements are well documented by the long list of references. This by itself is a major achievement, for which all those interested in the biosphere should be grateful.

In addition to presenting the broad overview, Smil indulges in a wealth of salient and often entertaining details. Did you know that a sperm whale can dive more than 2,000 metres deep, holding its breath for over an hour? Or that in 1973 a jet collided with a huge vulture at an altitude of over 11,000 metres? The book is replete with such *petites histoires*. These are welcome, of course, but point to a lack of organization underlying Smil's less systematic approach. At times I felt somewhat overwhelmed by the abundance of facts, observations, theories and details. This impression was strengthened by

GETTY IMAGES

the uneven style of writing. Some parts read like a newspaper, whereas others are in the dry style of professional reviews. I recommend several rounds of rigorous reorganization before the next edition goes into print.

With the present focus on Earth system science and geobiology, there is a pressing need for good introductory textbooks. Although I do not think that Smil's volume is sufficiently systematic to qualify for a first students' introduction to this burgeoning field, it will certainly be of great use to a more advanced audience with a general curiosity in biospheric matters. Smil conveys the thrill of exploring the unknown planet we inhabit. His enthusiasm and the broadness of his interest are infectious. ■

Peter Westbroek is emeritus professor in geology, Leiden University, 23344 AR Leiden, The Netherlands.

Putting scientists in the picture

Envisioning Science: The Design and Craft of the Science Image
by Felice Frankel

MIT Press: 2002. 335 pp. \$55, £36.50

Dee Breger

Newspaper mogul William Randolph Hearst had it right when he wrote: "The illustrations are a detail, though a very important one. Illustrations embellish a page; illustrations attract the eye and... materially aid the comprehension of an unaccustomed reader." Hearst was referring to artists' illustrations of newspaper stories in the late nineteenth century, but his insight was prescient about today's photography. He was later an avid photographer, so he would surely have been delighted at the burgeoning of photographic techniques that has occurred since the simple optics of his day. Technological developments have led to the atomic, microscopic, macroscopic and galactic scales of contemporary science images.

Even mediocre pictures can briefly attract the eye, if that's all we want from them. Better, though, is to enhance the viewer's comprehension by elevating images to media quality with the application of techniques used by professional artists and photographers. How to bring these techniques to the research community? Enter Felice Frankel and her exceptional manual, *Envisioning Science*.

Frankel, a photographer at the Massachusetts Institute of Technology, is a leader in the growing movement to enhance the quality of technical images from science and engineering. Her goal in this meticulously organized book is to help readers create "an image that communicates your work more

effectively to both colleagues and the general public". These are two distinct audiences, but Frankel's mission suggests that they don't necessarily require divergent approaches. A high-quality, technically and aesthetically proficient image is always an effective vehicle for transmitting information. A poorly lit or composed photo can convey less information than intended, and one that includes distracting extraneous details can obscure the message as well. These are especially important considerations for interdisciplinary communication.

The use of stimulating images attracts the often science-shy public as nothing else can; accompanying textual information is then more likely to be absorbed in a fluid, agreeable way. And, as Frankel points out, public interest creates a supportive climate for research. For all these reasons, universities have begun to recognize and invest in the powers of higher-quality scientific imagery.

Written in a conversational tone, *Envisioning Science* is primarily a comprehensive lab manual for the optimum photography of research samples. Covering both film-based

and digital techniques, it is also a generous reference book for photographic facts and resources, and includes exercises, a bibliography and two useful indexes. But it is also an art book in the sense that it is a beautifully designed array of arresting images (nearly all the author's). And it is a science guidebook in that the images chosen to illustrate the points under consideration derive from many disciplines and tell diverse stories.

An introductory section sets the stage for the core of the book: chapters that present concise, easy-to-assimilate descriptions of equipment and methods, copiously illustrated with comparative or before-and-after photographs. The initial chapter on the basics of picture-making includes data on sample preparation, composition, lighting, exposure and depth of field. Subsequent chapters tailor these topics to the photography of small samples 4–10 cm in size, and for using a stereomicroscope to document samples ranging from 50 μm to 4 mm. A chapter on compound microscopes (for samples down to 1 μm) includes a section on biological fluorescence microscopy. The final



Swirl of colour: ice can be seen melting in water by photographing gradients in the refractive index.

K. VANDIVER



Spice it up: images of gas deposition on a patterned polymer (top) can be made more compelling (bottom) with a bit of imagination.

chapter discusses archiving, scanning and communicating with the public, including an important comment on digital alteration of scientific images. The manual also covers the representation of motion and looks at related image-producing technologies.

Envisioning Science has everything that researchers need to start producing far higher-quality photographs of their own samples. The book's few minor problems are overwhelmed by an abundance of data, provided in succinct nuggets of information, encompassing the hows and whys of improving scientific photography for papers, conferences, journal covers and public venues. Non-science photographers could also benefit from this manual, as will general readers with an interest in science. ■

Dee Breger is at the Lamont-Doherty Earth Observatory, Palisades, New York 10964, USA.

Physician, reveal thyself

Doctors and Discoveries: Lives that Created Today's Medicine

by John Galbraith Simmons

Houghton Mifflin: 2002. 480 pp. \$24

W. F. Bynum

Encores can be tricky. Do you provide more of the same, or show your versatility by doing something completely different? John Simmons' *The Scientific 100* (reviewed in *Nature* 385, 215; 1997) audaciously but successfully offered a biographical ranking of the top scientists of all time. The encore, *Doctors and Discoveries*, which attempts something

similar for the medical profession, is geared to the author's strengths: an empathy for individuals and their contributions, and an outstanding ability to explain complex scientific and medical ideas simply yet accurately. There are differences, too. There are 15 fewer doses, only 85 entries, and the absolute ranking has been abandoned. Entries for individuals who appear in both volumes have been completely rewritten.

Happily, Simmons has not entirely given up the subjective dimension, which was part of the fun of the earlier volume. Instead, in six gatherings, he has stamped his judgement on the history of medicine. Rather than a simple chronological ordering, the result is a game in which the company of one's peers becomes significant.

The first part establishes the parameters of medical modernity, through two ancient thinkers, Hippocrates and Galen, and six nineteenth-century giants, with Charles Darwin occupying pride of place. Darwin's role as head of the class is both amusing (he was, after all, a would-be doctor who couldn't stand the sight of blood) and appropriate. Evolutionary biology underpins much of our knowledge of organisms in health and disease, and ought to have been more intertwined with biomedical thought from the start. The piece on Darwin takes account of recent developments in "evolutionary medicine", which promise much.

This pantheon is followed by a section on the highly commended, including several who were too clinical to make the first volume. DNA is nicely pinned on Oswald Avery, who showed that DNA is responsible

for heredity, instead of James Watson and Francis Crick, who were in the earlier volume. Besides, they will get plenty of exposure next year, the fiftieth anniversary of their famous *Nature* communication.

The third section, on historical also-worthies, records important characters such as Ambrois Paré, who discovered in the sixteenth century that it was better to treat wounds with a gentle salve rather than boiling oil; Pierre Fauchard, an eighteenth-century pioneer of dentistry; and psychoanalyst Sigmund Freud.

The next part spans the period from William Osler in the late nineteenth century to Macfarlane Burnet in the twentieth. This section has a Nobel prize count of about 75%, giving scant encouragement to those who might wish to improve diagnosis, medical education or treatment if their goal in life is the Ultimate Reward. Osler and the educationalist Abraham Flexner would not have expected one. Archibald Garrod died before the fundamental significance of his work on inborn errors of metabolism was adequately appreciated. Walter Cannon probably deserved one for his work on homeostasis and on the functions of the autonomic nervous system; neurosurgeons Harvey Cushing and Wilder Penfield would have liked one. It is not too late to consider Willem Kolff, who invented the kidney dialysis machine 60 years ago, because Peyton Rous, also included here, received his at the age of 87 for work on tumour-inducing viruses that he had done long before.

The Nobel count in the next, more recent, section is about the same, but it also includes

New in paperback

Voodoo Science: The Road from Foolishness to Fraud

by Robert L. Park

Oxford University Press, £8.99

"Park is a superb storyteller, he writes gracefully with energy and charm, and the varieties of silliness he chronicles are fascinating. Most scientists will enjoy the book immensely... *Voodoo Science* is a work that most scientists will applaud because of its wonderfully entertaining histories of deception and self-deception." N. David Mermin, *Nature* 407, 17–18 (2000).

Mendeleyev's Dream: The Quest for the Elements

by Paul Strathern

Berkley, \$14

"Although the book is full of detail, it always remains pleasantly readable... More can be gleaned from this informally written book, which is well recommended to the general reader." Gautam R. Desiraju, *Nature* 408, 29–30 (2000).

What Makes Us Think? A Neuroscientist and a Philosopher Argue About Ethics, Human Nature, and the Brain

by Jean-Pierre Changeux & Paul Ricoeur
Princeton University Press, \$17.95, £12.95

The Dragon Seekers: How an Extraordinary Circle of Fossilists Discovered the Dinosaurs and Paved the Way for Darwin

by Christopher McGowan
Perseus, \$17

Thieves, Deceivers, and Killers: Tales of Chemistry in Nature

by William Agosta
Princeton University Press, \$16.95, £11.95

The Madness of Adam and Eve: How Schizophrenia Shaped Humanity

by David Horrobin
Corgi, £8.99
"The book is vivid, sweeping and readable." Daniel Nettle, *Nature* 412, 119 (2001).

mavericks such as the paediatrician Benjamin Spock, sex physician William Masters and, curiously, Melanie Klein — curiously because psychoanalysis no longer has the psychiatric cogency that it once did. There is no mention of psychiatrist Emil Kraepelin, who could have been interestingly paired with his almost exact contemporary Freud. British readers might favour Richard Doll and Austin Bradford Hill ahead of Ernst Wynder on the smoking and health issue, especially as the rigour of Doll and Hill underpins epidemiological surveillance, double-blind clinical trials and, indirectly, evidence-based medicine. But Wynder was an influential figure in contemporary lifestyle medicine.

That fewer than half of the “creators” of

modern medicine worked in the United States, whereas two-thirds of the contemporary ones did (or do), is historically tenable. The only coupled entry is for HIV’s inextricably linked Robert Gallo and Luc Montagnier; Simmons’ exposition of their rivalry is both lucid and balanced.

I doubt that Simmons frequents either homoeopaths or chiropractors, but in his final section he discusses dispassionately some creators of modern medicine, if not exactly modern medical science. He points out that Avicenna, who died over a thousand years ago, is still studied widely, and that Samuel Hahnemann and D. D. Palmer, creators of homoeopathy and chiropractic, respectively, have many followers. Lydia Pinkham, a nineteenth-century American

nostrum pioneer, represents that happy band of entrepreneurs; Paul de Kruif, author of *Microbe Hunters* and other popular books, exemplifies an earlier heroic depiction of medicine; and Henri Dunant, founder of the Red Cross, provides a humanitarian face.

Medical schools have more difficulty recruiting than they did a generation ago. It is one of the many virtues of this book that it can be read and understood by someone thinking about a career in medicine. This young reader would be apprised of many of the difficulties facing anyone making such a career choice, but might still be encouraged to follow that path. ■

W. F. Bynum is at the Wellcome Trust Centre for the History of Medicine, 24 Eversholt Street, London NW1 1AD, UK.

Science in culture

Medal remodelled

Dutch artist Mirjam Mieras has designed the new Biochemical Society Award.

Martin Kemp

When we learn that a major scientific society is to award a medal to a star performer, we know roughly what to expect. Following Renaissance precedent, the medal will resemble a large coin, often with a portrait of a past luminary on the front and some kind of symbol or representative device on the reverse. It is likely to be made of bronze, or more expensively of silver or even gold. It will be suitably inscribed and come in a presentation case. The new biennial medal of the Biochemical Society shatters all these expectations.

The society, for whom I have been acting as adviser, have gone for a bold choice. They have commissioned the Amsterdam artist Mirjam Mieras to bring her innovative vision to bear on the task of designing something that pays homage to the essence of medallion design while departing from the stock conventions. The artist, for her part, has been determined to avoid the obvious clichés — most notably the ubiquitous double helix of DNA.

Mieras has established herself as one of the most innovative medallists in the world. She studied sculpture at the Royal Academy of Fine Arts in the Hague, and first made her mark as a medallist when she won the *Fédération Internationale de la Médaille* competition in the Hague in 1996. Her winning Moon and Mouse medal challenged the norm, using unusual materials — cloth, tin, glass and photographs — and making clever play around the circular motif, involving lenses, size and distance.

For the Biochemical Society she has produced a design that at first sight looks so minimal as to convey little. Once the recipient has slipped off the sleeve, reticently inscribed in shiny black on matt black, he or she is confronted with a heavy steel slab, relieved only by a translucent Plexiglas block at its centre. Once the block is slipped out,

a steel ring invitingly protrudes, courtesy of a hidden spring. When the ring is removed, we can see that its edge carries inscriptions. “RESEARCH RAISES QUESTIONS” runs around half of its rim, while the other half tells us that, “QUESTIONS GUIDE RESEARCH”, inverted in orientation. The letters A, T, C and G — for adenine, thymine, cytosine and guanine, the four bases of DNA — are highlighted in typical, textbook colours.

In a manner characteristic of much Dutch design, its apparent simplicity masks a complex set of subtle meanings and invites spectators to seek their own associations. Its precision-engineered, high-tech nature sits happily in the world of contemporary science. In dialogue with Keith Gull, then chairman of the society, she was directed to cassettes as a point of unexpected reference for the Plexiglas block. The symmetries, asymmetries, mirroring and inversions function particularly well in relation to the characteristics of many molecular structures. Such relationships are less the result of conscious adoption of scientific motifs, and more the consequence of intuitive affinities.

More broadly, Mieras finds that the circle, particularly in the context of its square container, presents possibilities that are at once bounded and limitless. Indeed, she confesses that she is “seduced by the circle”. As she says: “It took me a long time to understand the beauty and the rules of the circle. There is no beginning or end. A circle has no top or bottom. The top and bottom are determined by the observer. A medal is a three-dimensional circle. The optimum three-dimensional circle is a ball. The minimum three-dimensional circle is a pancake. All art medals lie somewhere between these two extremes. The possibilities are unlimited. I wonder if we are making use of all these possibilities. We must not lock up medal-making in tradition.”

Mieras’ stance, which respects the inherent rules of beauty of the genre with which she is

working while seeking radically new solutions, is more in keeping with the ethos of research science than the conventional medallion image of a grandee (generally male) in austere profile. *Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.*

The Biochemical Society Award 2002, designed by Mirjam Mieras, will be awarded to Bernard Dixon and Steven Rose for their contribution to science communication. Mieras’ medals will be exhibited at the Whitworth Art Gallery in Manchester, from 22 November to 22 December 2002.

Visualizations: The Nature Book of Art and Science is a collection of essays edited by Martin Kemp (published by Oxford University Press and the University of California Press; £20, £35).



Matter all in the mind

Kurt Gottfried

In the autumn of 1925, Paul Adrien Maurice Dirac, a Cambridge University graduate student of very few words, astonished the world of physics with a remarkable paper on quantum mechanics. His publication heralded the arrival of a mind of exceptional originality and power at the frontier of physics.

Werner Heisenberg's discovery of quantum mechanics sprang from the fertile soil of Göttingen and Copenhagen, not Cambridge. Nevertheless, Dirac, knowing only Heisenberg's ground-breaking but rather mysterious and fragmentary paper, produced an almost complete formulation of quantum mechanics wholly on his own, mirroring the achievement of the experienced Göttingen collaboration of Max Born, Pascual Jordan and Heisenberg. Born recalled this as "one of the greatest surprises of my scientific life, for the name of Dirac was completely unknown to me".

Dirac, born 100 years ago this year, often displayed an uncanny ability, whenever the need arose, to invent deep mathematical concepts that were new to physicists. Readers of his early paper would ask by what magic had he turned his austere, abstract constructs into quantitative descriptions of physical phenomena. Dirac's genius was quickly recognized — he was the youngest participant in the elite Solvay Congress of 1927, at which Niels Bohr and Albert Einstein began their long debate about the foundations of quantum mechanics. On first meeting him there, the formidable Wolfgang Pauli quipped that Dirac believed "there is no God, and Dirac is His prophet".

The development of nonrelativistic quantum mechanics — the microscopic counterpart of newtonian mechanics — was essentially complete by the end of 1926. But it was not known how to extend the new theory to the electromagnetic field, or to particles moving with velocities approaching the speed of light. Dirac solved both problems, giving birth to quantum electrodynamics.

Although Dirac's 'quantization' of classical electrodynamics yielded no great surprise, it was the first consistent and correct account of the absorption, emission and scattering of light. Ultimately, all the optical phenomena with which we are so familiar are fully described by Dirac's radiation theory of 1927.

By contrast, Dirac's relativistic wave equation for electrically charged particles, developed the following year, held many surprises — as he later put it, the equation is much more intelligent than its inventor. However one interprets this remark, the invention itself is an unsurpassed example of a disciplined imagination obtaining profound insights into the natural world.

Dirac's equation, when applied to the electron, describes accurately its magnetic moment, and elucidates fine details in the spectrum of hydrogen that are not accounted for by the nonrelativistic theory. Yet the equation seemed to have a terrible flaw, because it also had solutions with negative energy that made no apparent sense. Much effort was devoted to eliminating them, and when that failed, to understanding their significance.

After various inconsistent attempts by himself and others, Dirac made the spectacular proposal that these solutions describe a hitherto unknown particle that has the same mass but the opposite charge to the electron. Furthermore, he predicted that when this 'positron' encounters an electron, the pair would annihilate into two or more photons, and, conversely, that electron-positron pairs could be created by the interaction of energetic photons with matter. Soon after, the positron was discovered in cosmic-ray experiments, and was eventually found to have all the properties predicted by Dirac's theory.

Physics has produced other far-fetched predictions that have subsequently been confirmed by experiment. But Dirac's prediction of antimatter stands alone in being motivated solely by faith in pure theory, without any hint whatever from data, and yet revealing a deep and universal property of nature. The positron illustrates the general principle that all particles that have a conserved attribute, such as charge, have corresponding antiparticles. Furthermore, the existence of antimatter was soon understood to be an inevitable consequence of a

Quantum electrodynamics

Paul Dirac's prediction of antimatter stands alone in being motivated solely by faith in pure theory, without any hint whatever from data.

consistent marriage of relativity with quantum mechanics. A closely related lesson was that the particles that emerge from a reaction need not have existed beforehand — they can be created in the process itself, by the transformation of energy into matter. This fact has become so well established that it is now treated as mundane, but at the outset Dirac's theory was dismissed as incredible by Bohr and other leading figures, even as the evidence for the positron was accumulating.

Quantum electrodynamics has survived more searching experimental scrutiny than any other physical theory. It is valid down to distances more than six orders of magnitude shorter than those that had been explored at the time Dirac invented it. This accomplishment relies on great refinements in mathematical techniques, some of which, such as Richard Feynman's path integral, stem from ideas originating with Dirac. Today, quantum field theory, of which quantum electrodynamics is the prototypical core, is universally considered to be the most successful, fruitful description of the basic constituents of matter and their interactions.

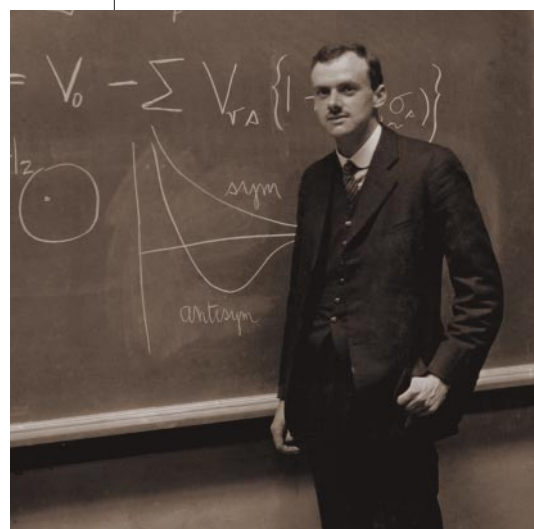
Dirac's creations are a constant presence in today's physics. His abstract, elegant style is now the standard language of quantum mechanics. His maxim that "physical laws should have mathematical beauty" guides throngs of physicists who pursue highly speculative research in the hope that they, too, can discover deep truths his way. The future may confirm another of Dirac's remarkable and brilliant ideas — that magnetic monopoles should exist, as this would explain why all particles carry an electrical charge of an integer multiple of the electronic charge. Indeed, Dirac monopoles, and closely related constructs, loom large in attempts to forge a unified theory of the fundamental interactions. ■

Kurt Gottfried is in the Laboratory of Nuclear Studies, Cornell University, Ithaca, New York 14853, USA.

FURTHER READING

Schweber, S. S. *QED and the Men Who Made It* (Princeton Univ. Press, 1994).
Kragh, H. *Dirac: A Scientific Biography* (Cambridge Univ. Press, 1990).
Miller, A. I. *Early Quantum Electrodynamics* (Cambridge Univ. Press, 1994).

AIP/SPL



Truth is beauty: Paul Dirac and the quantum mechanics of the hydrogen molecule.

Casimir force changes sign

Eyal Buks and Michael L. Roukes

This quantum attractive force induces measurable effects between ultrasmall mechanical components. New calculations indicate that systems could be engineered in which Casimir forces are repulsive.

In 1948, Hendrik Casimir calculated that the quantum fluctuations of an electromagnetic field, so-called zero-point fluctuations, give rise to an attractive force between objects¹. This force is a particularly striking consequence of the quantum theory of electrodynamics (for a review, see ref. 2). Casimir's calculations were idealized — he considered two perfectly conducting parallel plates at absolute-zero temperature — but there are implications for more realistic objects. In *Physical Review Letters*, Kenneth *et al.*³ have extended these considerations to real-world materials.

Their work follows that of Boyer in 1974, who also studied the case of parallel plates but with one plate perfectly conducting and the other having infinite magnetic permeability (permeability is a measure of the material's response to an applied magnetic field). For this special case Boyer found that quantum fluctuations induce a force with the opposite sign, causing the plates to repel each other⁴. Kenneth *et al.* now extend understanding of the Casimir force phenomenon to the more general situation of realistic 'dielectric' materials that are characterized by both their electrical permittivity (a measure of the material's response to an applied electric field) and their magnetic permeability. Their numerical results show that repulsive forces can arise in the general class of materials with high magnetic permeability.

Although the Casimir effect is deeply rooted in the quantum theory of electrodynamics, there are analogous effects in classical physics. A striking example was discussed in 1836, in P. C. Caussee's *L'Album du Marin* (*The Album of the Mariner*)⁵. Caussee reported a mysteriously strong attractive force that can arise between two ships floating side by side — a force that can lead to disastrous consequences (Fig. 1). A physical explanation for this force was, however, offered only recently, by Boersma⁶, who suggested that it originates in the radiation pressure of water waves acting differently on the opposite sides of the ships.

His argument goes as follows: the spectrum of possible wave modes around the two ships forms a continuum (any arbitrary wave-vector is allowed); but between the vessels their opposing sides impose boundary conditions on the wave modes, restricting the allowed values of the component of the wave-vector that is normal to the ships'

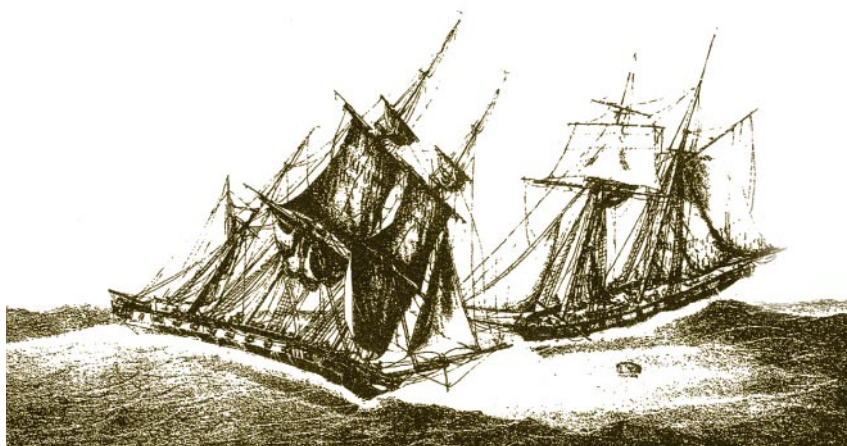


Figure 1 A Casimir-like effect at sea. In the days of square-riggers, sailors noticed that, under certain conditions, ships lying close to one another would be mysteriously drawn together, with various unhappy outcomes. Only in the 1990s was the phenomenon explained as a maritime analogy of the Casimir force. (Illustration from ref. 5.)

surfaces. This discreteness created in the spectrum of wave modes results in a local redistribution of modes in the region between the ships, with the consequence that there is a smaller radiation pressure between the ships than outside them.

Analogous arguments had previously been employed by Milonni *et al.*⁷ to explain the origin of the Casimir effect itself. In this case, the radiation pressure is due to electromagnetic waves rather than water waves. Casimir had considered a system at zero temperature in which, classically, no radiation pressure is expected. But the quantum theory of electrodynamics states that the electromagnetic field exhibits quantum fluctuations even at zero temperature, and these are the source of Casimir forces acting on macroscopic bodies. Another outcome of these quantum fluctuations is the van der Waals force⁸, which, in essence, can be considered as the Casimir force at especially small separations.

Historically, the Casimir effect has been considered to be an exotic quantum phenomenon, but now it is starting to take on technological importance. Because of its relatively short range, it has only a very small effect on the dynamics of macroscopic mechanical systems. But the Casimir force has a major role in modern micro- and

nanoelectromechanical systems (MEMS and NEMS), where the distances between neighbouring surfaces are typically far less than 1 μm (ref. 9). This new branch of microelectronics uses the methodology of integrated-circuits manufacturing for the fabrication of on-chip, fully integrated, miniature sensors and actuators, with a rapidly growing range of applications.

In tiny devices such as these, the Casimir force can cause mechanical elements to collapse onto nearby surfaces, resulting in permanent adhesion — an effect called 'stiction', which often proves to be an important factor in the malfunction of NEMS. Follow-up theoretical and experimental studies to the work of Kenneth *et al.*³ might uncover ways to engineer NEMS in which the Casimir forces are repulsive. They may even open the way for new applications of NEMS that are, in effect, immune to stiction.

Kenneth *et al.* also emphasize that the Casimir force is non-additive. For additive forces, the total force acting on a body is simply the sum of the pairwise contributions between bodies — for example, the total electrostatic force on an element A, interacting with elements B and C, is found by adding the force between elements B and A and the force between elements C and A. For

the Casimir force, however, the additive approach can break down completely. For the case they considered, Kenneth *et al.* show that the additive pairwise approach can even erroneously predict attractive forces, although the exact calculation proves that the forces are repulsive.

The development of vital tools, such as scanning probe microscopy and MEMS and NEMS technology, has made possible a new generation of experiments², and the importance of the Casimir phenomenon for both fundamental physics and practical applications is now becoming more widely appreciated. The interplay between basic science and technology is certain to motivate the study of Casimir forces in the years to come. ■

Eyal Buks is in the Department of Electrical

Engineering, Technion, Haifa 32000, Israel.

e-mail: eyal@ee.technion.ac.il

Michael L. Roukes is in the Condensed Matter Physics Group, Caltech 114-36, Pasadena, California 91125, USA.

e-mail: roukes@caltech.edu

1. Casimir, H. B. G. *Proc. Kon. Ned. Akad.* **51**, 793–795 (1948).
2. Bordag, M., Mohideen, U. & Mostepanenko, V. M. *Phys. Rep.* **353**, 1–205 (2001).
3. Kenneth, O., Klich, I., Mann, A. & Rezen, M. *Phys. Rev. Lett.* **89**, 033001 (2002).
4. Boyer, T. H. *Phys. Rev. A* **9**, 2078–2084 (1974).
5. Caussée, P. C. *L'Album du Marin* (Mantes, Charpentier, 1836).
6. Boersma, S. L. *Am. J. Phys.* **64**, 539–541 (1996).
7. Milonni, P. W., Cook, R. J. & Goggin, M. E. *Phys. Rev. A* **38**, 1621–1623 (1988).
8. Milonni, P. W. *The Quantum Vacuum – An Introduction to Quantum Electrodynamics* (Academic, New York, 1994).
9. Chan, H. B., Aksyuk, V. A., Kleiman, R. N., Bishop, D. J. & Capasso, F. *Science* **291**, 1941–1944 (2001).

DNA repair

Right on target with ubiquitin

Cecile M. Pickart

Cellular DNA-repair mechanisms prevent mutations from accumulating, thereby averting defects in cell function. A molecule best known for its role in protein degradation is now shown to have a specific task in DNA repair.

The activity of many of the proteins in our cells depends on the chemical ‘labels’ that are attached to them. A commonly used label is the ubiquitin molecule, which regulates numerous cellular processes¹. Many of ubiquitin’s regulatory functions reflect its role as a tag that singles out proteins to be degraded¹. Yet ‘ubiquitination’ can also signal other fates². The idea that cells can interpret ubiquitin signals in diverse ways was first suggested more than a decade ago, when a protein called Rad6 — which is needed for cells to repair damaged DNA³ — was found to be an enzyme that helps to join ubiquitin to target proteins⁴. Despite this early hint, further evidence of a role for ubiquitin in protecting DNA was elusive: the relevant target proteins, and the consequences of their modification, remained unknown. But, on page 135 of this issue, Hoege and co-workers⁵ identify the first functionally relevant target of ubiquitination during DNA repair — a protein called PCNA.

DNA is highly susceptible to environmental insults that can alter its sequence (causing mutations), or prevent it from being copied altogether (causing cell death)⁶. For instance, DNA damage caused by chemicals or ultraviolet light can block the progress of the enzymes that copy DNA — DNA polymerases — creating gaps in one of the two strands of a newly produced DNA molecule. Ubiquitination is a positive signal in a pathway that permits DNA replication to be completed despite such damage⁷. This pathway, which requires the Rad6 protein, does not remove the original lesions, but instead uses ‘post-

replicative’ DNA synthesis to fill in the gaps, using an undamaged strand as a template. The synthesis can follow either of two routes, one error-prone and the other error-free. The error-free mechanism is especially important — if it fails, the error-prone mechanism takes over, generating mutations in the newly made DNA. Post-replicative repair is the least well understood of the several DNA-repair mechanisms in higher organisms⁶.

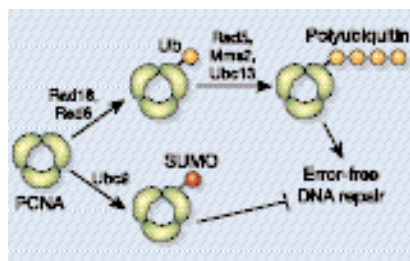


Figure 1 Connecting DNA repair and ubiquitin, through the PCNA protein. The PCNA trimer, shown at the left, encircles DNA and binds to DNA-replicating enzymes (polymerases). Hoege *et al.*⁵ have found that one complex of ubiquitin-conjugating proteins (Rad18 and Rad6) attaches a single ubiquitin (Ub) to a specific lysine amino acid in PCNA. A second conjugating complex (consisting of Rad5, Mms2 and Ubc13) extends a polyubiquitin chain from this first ubiquitin. The modified PCNA then promotes error-free post-replicative DNA repair. Modification of this same lysine amino acid by SUMO, another member of the ubiquitin family, depends on a distinct conjugating factor (Ubc9) and inhibits such DNA repair.

The main type of error-free post-replicative repair requires five enzymes, including Rad6, that conjugate ubiquitin to target proteins (Fig. 1)⁷. The pathway also relies on a specialized signal in which several ubiquitins are chained together in a particular way⁸. These properties bespeak an intimate involvement of ubiquitin in DNA repair, but several questions remained unanswered until now. Which proteins are modified with ubiquitin by the dedicated conjugating enzymes? Which polymerase (or polymerases) carries out the bulk of error-free post-replicative synthesis? And what is the consequence of labelling the target proteins with ubiquitin? In providing an answer to the first question — that PCNA is ubiquitinated during post-replicative repair — the work of Hoege *et al.*⁵ sets the stage for answering the second and third.

PCNA was already known to work with one DNA polymerase in DNA replication⁹, and with others in DNA repair. How did Hoege *et al.* discover that PCNA is also a target for ubiquitination? As is often the case, serendipity was important. The researchers actually set out to study SUMO, one of several ubiquitin-like proteins (the name stands for ‘small ubiquitin-related modifier’) that act through a similar conjugation mechanism¹⁰. Hoping to gain insight into SUMO-mediated signalling, the authors purified ‘sumoylated’ target proteins from yeast cells, and found that PCNA was among them.

While confirming this finding, Hoege *et al.* also noticed other modified forms of PCNA, which proved to be attached to the special chain of ubiquitins that is the signature of error-free repair. This ‘polyubiquitination’, but not sumoylation, required the Rad6 pathway to be active and was induced by DNA damage. Moreover, DNA repair was inhibited when the authors mutated the lysine amino acid in PCNA to which polyubiquitin becomes attached, and genetic analysis showed that this effect resulted from blocking the conjugation of PCNA to ubiquitin. So PCNA is truly a repair-relevant substrate for ubiquitination. But the effect on DNA repair of mutating PCNA’s ubiquitin-attachment site is weaker than the effect of blocking ubiquitin conjugation altogether, implying that further targets of polyubiquitination remain to be discovered.

So the polyubiquitination of PCNA is needed for DNA repair. What about the sumoylation of this protein? Ubiquitin and SUMO are attached to target proteins by different conjugating factors. Nonetheless, they can modify the same lysine residue of PCNA. Hoege *et al.* present several lines of evidence that suggest that occupation of this lysine by SUMO inhibits DNA repair. Such antagonism between ubiquitin and SUMO was known in protein degradation¹¹, but it seems that the strategy may apply more broadly.

With a known substrate in hand, Hoege

and co-workers were also able to gain insight into the specific roles of the various ubiquitin-conjugation enzymes. Apparently, a single ubiquitin attached to the crucial lysine of PCNA by one conjugating complex serves as the starting point for extension of the chain by a different complex. This persuasive model agrees with the order of events suggested by earlier genetic analyses and is consistent with several known protein-protein interactions^{7,12}. The production of a polyubiquitin chain by the sequential action of different conjugating factors is a new observation; it will be interesting to see which other ubiquitin-conjugating systems use this strategy.

PCNA had been implicated previously in post-replicative repair¹³ and is also involved in other repair pathways⁹. The results of Hoege *et al.* confirm the relevance of PCNA to Rad6-mediated repair, and lend weight to the idea that the DNA polymerase with which PCNA works in DNA replication is also intimately involved in error-free repair¹⁴. PCNA assembles into a trimeric ring that encircles DNA and promotes faithful and efficient replication by DNA polymerases. It also provides a platform for recruitment of regulatory factors⁹. The lysine residue modified by SUMO or polyubiquitin lies at the edge of the PCNA ring. One can easily imagine the polyubiquitin chain recruiting new molecules to a stalled polymerase, promoting its removal, relocalization or reprogramming. So the new results are the first to suggest how repair and replication could be integrated at the molecular level.

The results also raise new questions — for example, is the switch between sumoylation and ubiquitination controlled through attachment exclusively, or is the removal of SUMO and ubiquitin regulated too? And does attachment of a single ubiquitin (rather than a chain) to PCNA promote a different mode of repair, as the authors propose⁵? Finally, two of the ubiquitin-conjugating factors involved have additional biochemical activities: Rad18 binds single-stranded DNA and Rad5 may unwind DNA⁷. Whether these properties are relevant in interpreting the ubiquitin signal, as well as in generating it, remains to be seen. ■

Cecile M. Pickart is in the Department of Biochemistry and Molecular Biology, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, Maryland 21205, USA.

e-mail: cpickart@jhmi.edu

1. Hershko, A. & Ciechanover, A. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
2. Pickart, C. M. *Trends Biochem. Sci.* **25**, 544–548 (2000).
3. Reynolds, P., Weber, S. & Prakash, L. *Proc. Natl Acad. Sci. USA* **82**, 168–172 (1985).
4. Jentsch, S., McGrath, J. P. & Varshavsky, A. *Nature* **329**, 131–134 (1987).
5. Hoege, C., Pfander, B., Moldovan, G.-L., Pyrowolakis, G. & Jentsch, S. *Nature* **419**, 135–141 (2002).
6. Hoeijmakers, J. H. J. *Nature* **411**, 366–374 (2001).
7. Broomfield, S., Hryciw, T. & Xiao, W. *Mutat. Res.*

- 486, 167–184 (2001).
8. Spence, J., Sadis, S., Haas, A. L. & Finley, D. *Mol. Cell. Biol.* **15**, 1265–1273 (1995).
9. Warbrick, E. *BioEssays* **22**, 997–1006 (2000).
10. Hochstrasser, M. *Nature Cell Biol.* **2**, E153–E157 (2000).
11. Desterro, J. M., Rodriguez, M. S. & Hay, R. T. *Mol. Cell* **2**, 233–239 (1998).

12. Ulrich, H. D. & Jentsch, S. *EMBO J.* **19**, 3388–3397 (2000).
13. Torres-Ramos, C., Yoder, B. L., Bugers, P. M. J., Prakash, S. & Prakash, L. *Proc. Natl Acad. Sci. USA* **93**, 9676–9681 (1996).
14. Torres-Ramos, C. A., Prakash, S. & Prakash, L. *J. Biol. Chem.* **272**, 25445–25448 (1997).

Astronomy

The missing link

Shri Kulkarni

Some stars might be 'magnetars', powered by magnetism instead of fusion. The discovery of an X-ray burst from an anomalous X-ray pulsar suggests that this type of star can be added to the list.

About 60 years ago, physicists realized that ordinary stars are powered by nuclear fusion. Since then, astronomers have found other types of star that are powered by gravitational binding energy, by radioactivity or by rotational energy. Over the past decade, there has been speculation that stars powered entirely by magnetism — 'magnetars' — could also exist. Magnetism is quite common in the Universe. The Earth has a magnetic field of 1 gauss (G), and the Sun of 10 G. For comparison, a typical toy magnet on a refrigerator door has a strength of about 100 G, and the strongest sustained magnetic fields produced in laboratories are around 10^6 G. But magnetars are thought to have magnetic fields a billion times stronger still. On page 142 of this issue, Gavril, Kaspi and Woods¹ present evidence that an anomalous star 15,000 light years from Earth is a magnetar.

There is already one class of astronomical objects that astronomers think are mag-

netars^{2,3}. These are the 'soft-gamma-ray repeaters' (SGRs), which emit intense bursts of low-energy gamma-rays. One such burst was so powerful it caused detectable changes in the Earth's ionosphere. Initially confused with extragalactic gamma-ray bursts, SGRs are believed to be a subclass of the young neutron stars within our Galaxy⁴. The bursts are so intense that they can only be explained by the existence of strong magnetic fields confining the electrons that are radiating gamma-rays (similar to the magnetic confinement of plasma in a fusion reactor). The discovery of regular pulsations with long periods⁵ added further support for the existence of intense magnetic fields in SGRs.

Whether another subclass of neutron stars are magnetars has been more controversial. These are a group of X-ray pulsars with similarly long periods — the so-called 'anomalous X-ray pulsars' (AXPs). Pulsars are thought to be spinning neutron stars, and

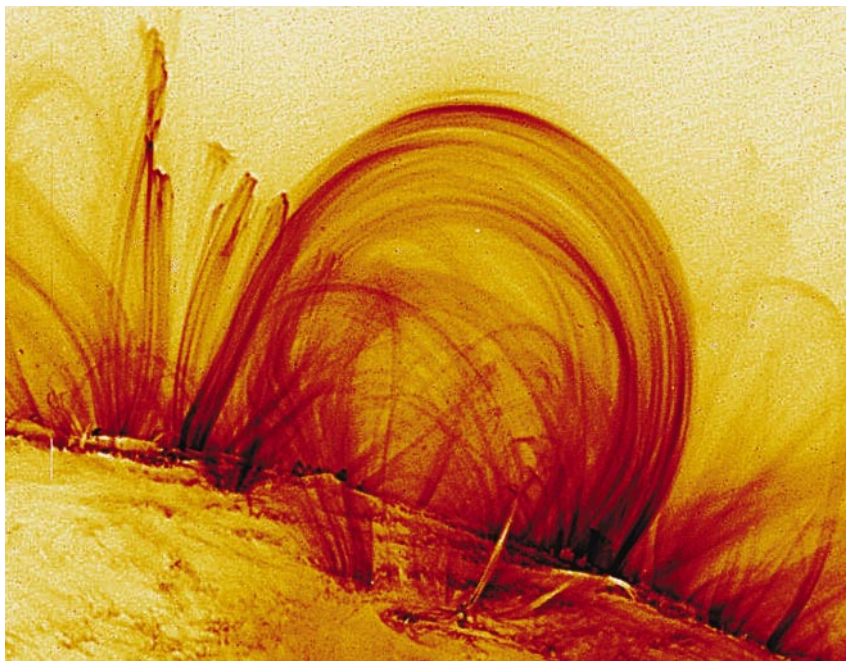


Figure 1 The surface of the Sun photographed by the TRACE spacecraft. Hot ionized gas forms coronal loops, tracing out the magnetic-field structure at the star's surface.

NASA/TRACE

ordinary X-ray pulsars are powered by the matter that they accrete from a companion star. But, despite careful searches, no plausible companions have been identified around AXPs. Nonetheless, AXPs show persistent and strong X-ray emission, well above the level that can be supported by their store of rotational energy. The source of energy in these objects is mysterious, which is why these X-ray pulsars are called anomalous. The long rotation period of AXPs also suggests a family resemblance to SGRs.

According to the magnetar model devised by Thompson and Duncan⁶, the decay of the intense magnetic field in a magnetar provides a source of energy and powers the radiation emission. By analogy with the Sun (Fig. 1), loops of magnetic field could occasionally reconnect, producing a flare. But in the case of magnetars, the field strength is 10^{12} times greater than that in the loops on the Sun, and the resulting flares are proportionally more intense. Furthermore, the strong magnetic field might move portions of the outer layer, or crust, of the neutron star — as though moving continental plates. Like earthquakes, a major flare could be followed by minor ones. Once the stress has built up once more, the reconnection process happens all over again, which explains the repetitive appearance of the flares.

There are, however, several issues still outstanding for the magnetar model. To start with, there is the problem of the evidence for strong magnetic fields in SGRs. From the spectra of emitted light, astronomers have been able to measure field strengths directly in a number of astrophysical objects, including the Sun and ordinary neutron stars. Yet there had been no such direct spectral evidence for SGRs: evidence for their strong magnetic fields came only indirectly, through their intense bursts of radiation. Second, the magnetar model postulates that both AXPs and SGRs are magnetars, and yet the two types of star seem quite different in that no bursts of radiation had ever been detected from AXPs. So although the case for SGRs as magnetars is plausible, for AXPs it has up to now been considerably weaker.

But during a routine monitoring programme using the Rossi X-ray Timing Explorer, Gavril *et al.*¹ found bursts from an AXP. They argue quite persuasively that these bursts arise from the AXP and not from an unrelated source. This discovery at last establishes a strong link between AXPs and SGRs.

There is a plausible explanation for why the bursts from AXRs are fainter than those from SGRs, and thus harder to detect. AXPs tend to be located at the centre of supernova remnants, and so are undoubtedly young neutron stars. In contrast, SGRs are not associated with supernova remnants, implying that SGRs must be older than AXPs⁷. It is possible that the crust of young neutron stars is more malleable or plastic, and thus AXPs may

be unable to support superstrong magnetic loops that reconnect to generate the intense bursts. But as AXPs age, the crust could become less plastic and consequently be capable of supporting strong magnetic loops — up to a point.

Gavril *et al.*¹ also made spectroscopic measurements. They found that the bursts have a strong feature in their spectrum that can reasonably be interpreted as being due to protons gyrating in the intense fields of a magnetar. A similar spectral feature has recently been reported for an SGR⁸. Both detections await firm confirmation — and will motivate many further observations — but if they are real then we have direct evidence for strong field strengths.

It seems that magnetars do exist. Astronomers can feel quite satisfied to have postulated, discovered and confirmed a new class of cosmic object. And physicists will be excited by the possibilities of magnetars:

where magnetic field strengths exceed about 10^{14} G, all sorts of strange effects arising from the quantum theory of electrodynamics are potentially detectable (for example, the intense fields of magnetars could polarize the vacuum). Sensitive experiments could look for such effects and make good use of one of nature's greatest laboratories. ■

Shri Kulkarni is in the Department of Physics, Mathematics and Astronomy, Caltech, Pasadena, California 91125, USA.

e-mail: srk@astro.caltech.edu

1. Gavril, F. P., Kaspi, V. M. & Woods, P. M. *Nature* **419**, 142–144 (2002).
2. Thompson, C. & Duncan, R. C. *Astrophys. J.* **408**, 194–217 (1993).
3. Pacynski, B. *Acta Astron.* **42**, 145–153 (1992).
4. Kulkarni, S. R. & Frail, D. A. *Nature* **365**, 33–35 (1993).
5. Kouveliotou, C. *et al.* *Nature* **393**, 235–237 (1998).
6. Thompson, C. & Duncan, R. C. *Mon. Not. Astron. Soc.* **275**, 255–300 (1995).
7. Gaensler, B. M., Slane, P. O., Gotthelf, E. V. & Visishit, G. *Astrophys. J.* **559**, 963–972 (2001).
8. Ibrahim, A. I. *et al.* *Astrophys. J.* **574**, L51–L55 (2002).

Global change

Oceanic action at a distance

Raja S. Ganeshram

Glacial intervals are characterized by low levels of atmospheric CO₂. A new explanation for that connection invokes nutrient export from the Southern Ocean to warmer waters at such times.

In the late 1980s came news of a spectacular discovery: that temperatures at Earth's surface and levels of CO₂ in the atmosphere fell and rose in tandem during glacial–interglacial cycles¹. Since then, Earth scientists have been busy seeking explanations for the connection and much recent thinking has centred on oceanic controls on CO₂. Writing in *Geophysical Research Letters*² and *Global Biogeochemical Cycles*³, respectively, Brzezinski *et al.* and Matsumoto *et al.* take this line of investigation a step forward. Together the papers provide a new hypothesis, elegantly supported by a set of palaeoceanographic evidence and biogeochemical modelling, that could help account for much of the reduction in atmospheric CO₂ during glacial periods.

For some years, Milankovitch cycles — minute variations in solar radiation reaching the Earth due to changes in Earth's orbit — were taken as an explanation for the pacing of the oscillations between glacial and interglacial intervals. But accounting for the large temperature shifts involved required further amplification within the Earth's climate system. Once analyses of air trapped in ice cores¹ revealed that CO₂ had fallen and risen in tune with the glacial–interglacial rhythm, the finger pointed to the greenhouse effect exerted by changes in CO₂ as the amplifier.

Certain parts of the world's oceans have an especially large influence on climate, the Southern Ocean — the waters around

Antarctica south of the polar front — being one of them. Oceanographers have long recognized that reduced leakage of CO₂ from ocean to atmosphere in such areas could contribute significantly to the lower levels of atmospheric CO₂ in glacial periods. In today's Southern Ocean, CO₂-charged, nutrient-rich deep waters reach the surface and release CO₂ to the atmosphere. This leak could effectively be plugged by an increase in biological production in the surface waters through fuller use of the available nutrients by phytoplankton, primarily diatoms⁴. Increased production would transfer more CO₂ back to the ocean depths in the form of sinking organic particles.

Today, nutrients — particularly nitrate — that reach the sunlit surface ocean around Antarctica remain underused by phytoplankton because of a lack of the iron needed by their photosynthetic apparatus⁵. But during glacial periods, the larger amounts of dust in the atmosphere should have increased the supply of iron, potentially alleviating the deficiency, increasing nutrient use and yielding a net transfer of CO₂ back to deep waters. Isolating the deep waters from the atmosphere by capping them with less-dense water at the surface (stratification) would have the same effect. With these ideas in mind, palaeoceanographers have been studying Southern Ocean sediments laid down in glacial times for clues to higher

Although the terms "ass" and, at any rate in Germany, "ox" (Ochs) are very generally applied to stupid persons, those who have observed the bovine and asinine genera know that this is an injustice to those animals.... A donkey that was kept here learnt to open, not only the gate of its own field, but other gates. One day, having left its own abode, accompanied by two ponies, it went to another field half a mile off, opening three gates on the way, liberated the occupants of this field, a mare and her foal, and a yearling, old friends of the donkey's, as they used to live together, and the whole party, which had been joined by a mastiff, proceeded to wander through the world. About two miles from here the horses were recognised and secured, and the donkey eventually returned with the mastiff; but after this exploit it was thought advisable to get rid of the donkey, as being too zealously devoted to the cause of emancipation. From *Nature* 11 September 1902.

Scientific Progress of April (40, No. 158, 193; 1952) contains an interesting and informative article by Prof. H. S. W. Massey entitled "Fundamental Particles", the term which is applied usually to such entities as electrons, protons, etc. ... Turning next to the classification of the fundamental wave particles, Prof. Massey shows that the particles can be loosely separated into three categories, which he aptly terms "building stones" (electron, proton and neutron), "cements" (photon and pi-meson) and "bric-a-brac" (neutrino, mu-, V-, tau- and kappa-mesons). The first two categories, as their names imply, are involved in the structure of matter, but the third, apparently, does not fulfil any important role in that respect. The properties of the various particles and the relations between them are briefly and clearly described, and it is evident that their number (now some twenty-four) is far too large for them all to be fundamental. Nevertheless, as Prof. Massey states, the discovery of new particles is still a prominent feature of modern physics, and thus, until some new fundamental advance or simplification is made on the theoretical side, not only to provide a basis for the "bric-a-brac" but possibly also to account for the more complete range of particles yet to be explored, the fundamental scheme of Nature must remain obscure.

From *Nature* 13 September 1952.

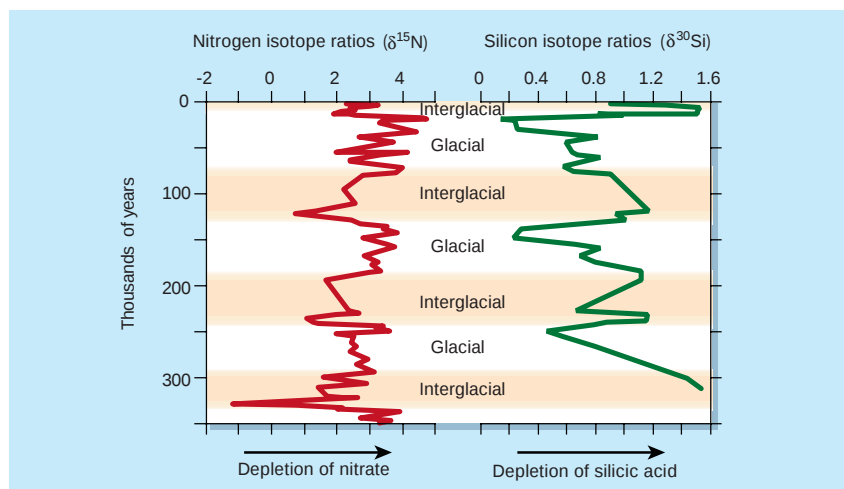


Figure 1 Variation of nitrogen and silicon isotopes from Antarctic sediments over the past 350,000 years. The signatures show opposite trends during glacial and interglacial periods. This pattern can be explained²³ by the influence of changes in iron availability on the ratio of nitrate and silicic acid used by diatoms in the Southern Ocean.

Stable isotope ratios of nitrogen and silicon in organic and diatom remains, respectively, tell us about the nutrient status during the geological past. This is because diatoms' use of nitrate (NO_3^-) and silicic acid ($\text{Si}(\text{OH})_4$) favours the uptake of the lighter isotopes, ^{14}N and ^{29}Si . Diatoms become progressively enriched in the heavier isotopes, ^{15}N and ^{30}Si , as the nutrients are depleted, and increased nutrient use should be reflected in their sedimentary remains as higher $\delta^{15}\text{N}$ and $\delta^{30}\text{Si}$. As shown in Fig. 1, sediment cores from the Southern Ocean have higher $\delta^{15}\text{N}$ values in the glacial intervals, indicating increased NO_3^- uptake compared to that in interglacials. But the $\delta^{30}\text{Si}$ trends are quite the opposite: $\text{Si}(\text{OH})_4$ use appears to have been lower during the glacials. Other arguments that invoke, for example, increased ocean stratification, could account for the higher $\delta^{15}\text{N}$ but they fail to explain lower $\delta^{30}\text{Si}$ values⁶. How can we reconcile the two different pictures of nutrient status painted by these two forms of proxy data?

On the basis of experimental results, Brzezinski *et al.*² point out that the addition of iron dramatically alters the uptake ratio of NO_3^- and $\text{Si}(\text{OH})_4$ by diatoms from as much as 4:1 to about 1:1. Given that ratios of these nutrients in the Southern Ocean today are 2:1, uptake with a ratio of 1:1 under iron-replete conditions, as might have occurred during glacial periods, would result in higher use of NO_3^- but relative underutilization of $\text{Si}(\text{OH})_4$ — a view that is consistent with the isotope data. Given this revised interpretation of the sediment records, did the Southern Ocean contribute to the drop in atmospheric CO_2 during glacial periods? Brzezinski *et al.*² argue that it did, but not as a direct effect of CO_2 uptake in Antarctic

The waters of the modern Southern Ocean penetrate north as far as the subtropics, mainly as subsurface flow. If the flow paths were the same during glacials, then the Antarctic region would have supplied the low-latitude ocean with water high in $\text{Si}(\text{OH})_4$ and low in NO_3^- . This would have pushed the phytoplankton community in these regions from domination by CaCO_3 -secreting coccolithophores to domination by opal-secreting diatoms. Brzezinski *et al.* believe, and Matsumoto *et al.* demonstrate with their model, that such an ecological shift would have had a double effect on reducing glacial CO_2 . First, diatoms have higher sinking rates than coccolithophores: their dominance would have resulted in the more efficient export of particulate organic matter to the depths, thus removing CO_2 from the surface waters. Second, the lowered ratios of CaCO_3 to organic carbon in sinking particles would also have lowered levels of atmospheric CO_2 , because the resulting excess CO_3^- (alkalinity) in surface waters would have sequestered CO_2 by converting it to bicarbonate.

But did such a shift towards increased primary production by diatoms actually occur at low latitudes in glacial times? The answer is unclear, because the telling indicator, accumulation of opal in sediments of glacial age, is not uniformly high in these regions. Increases are indeed seen in some areas⁷. Declines are evident elsewhere, however, such as in the eastern tropical Pacific⁸. This geographical variability may point to local changes in nutrient inputs as the main determinant of opal production.

The work of Brzezinski *et al.*² and Mat-

sumoto *et al.*³ resolves the conflict generated by the proxy evidence for the nutrient status of the Southern Ocean at various times. But assessing the effects of their proposed mechanism on atmospheric levels of CO₂ during glacials, as well as the effects of other ideas that have been put forward^{9,10}, will require more evidence. We need a more synoptic view of past shifts in nutrient ratios, and a better understanding of the ecological response to such changes. Progress is being made, but the story isn't over yet. ■

Raja S. Ganeshram is in the Department of Geology and Geophysics, University of Edinburgh,

West Mains Road, Edinburgh EH9 3JW, UK.

e-mail: Raja.Ganeshram@glg.ed.ac.uk

1. Barnola, J. M. *et al.* *Nature* **329**, 408–414 (1987).
2. Brzezinski, M. A. *et al.* *Geophys. Res. Lett.* **29**, 10.1029/2001GL014349 (2002).
3. Matsumoto, M., Sarmiento, J. L. & Brzezinski, M. A. *Glob. Biogeochem. Cycles* **16**, 10.1029/2001GB001442 (2002).
4. Sarmiento, J. S. & Toggweiler, J. R. *Nature* **308**, 621–624 (1984).
5. Boyd, P. W. *et al.* *Nature* **407**, 695–702 (2000).
6. Francois, R. *et al.* *Nature* **389**, 929–935 (1997).
7. Broecker, W. S. *et al.* *Paleoceanography* **15**, 348–352 (2000).
8. Ganeshram, R. S. & Pedersen, T. F. *Paleoceanography* **13**, 634–645 (2000).
9. Ganeshram, R. S. *et al.* *Nature* **415**, 156–159 (2002).
10. Devol, A. H. *Nature* **415**, 131–132 (2002).

Optics

The light fantastic

Martin Hegner

Optical tweezers use light to manipulate tiny particles — but only one at a time. If the light in the tweezers is a 'Bessel beam', this problem can be overcome, creating some interesting experimental possibilities.

Optical tweezers have become a standard tool in many areas of science, such as colloid research and biological studies. On page 145 of this issue, Garcés-Chávez *et al.*¹ report an extension of this technique that makes it possible to manipulate ensembles of particles simultaneously.

Laser light can exert force on dielectric (polarizable) particles through radiation pressure and refraction²: light has momentum, and so light and matter can interact by exchanging momentum. Although these forces are small (of the order of 10^{−12} newtons), they are sufficient to trap and manipulate micrometre-sized particles in a liquid environment³. So far, optical manipulation of particles in different places at the same time has been possible only if the particles are in the same optical plane of view⁴. Otherwise, the problem is that the manipulating light beam will diverge around the first trapped particles, and then cannot be brought back to a focus within the short distance to the next set of particles. So experiments must be done in series, and this is very time-consuming. To speed things up, it would be better to work in parallel, manipulating particles held in multiple back-to-back compartments (such as a collection of biological-cell samples) at the same time with a single light beam.

Using a special kind of optical focusing to create a 'Bessel beam' of light, Garcés-Chávez *et al.*¹ now provide the first demonstration of the manipulation of micrometre-sized particles in multiple planes. A Bessel beam (Fig. 1) is so called because the variation of its intensity follows the mathematical pattern known as a zero-order Bessel function. Thanks to its special optical properties, the bright central spot of the beam can re-form after passing an obstruction; the beam does

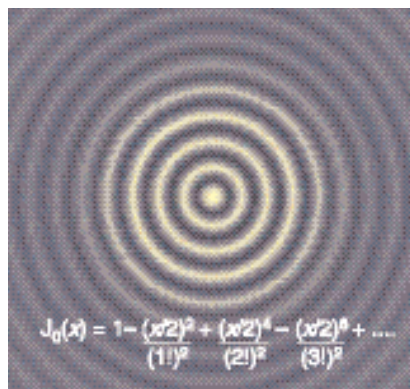


Figure 1 A Bessel beam. Light passing through a particular scheme of lenses can form a beam of light with a bright central spot and concentric rings of decreasing intensity — known as a Bessel beam, as its varying intensity is described by the zero-order Bessel function shown here. Garcés-Chávez *et al.*¹ have used a Bessel beam to create optical tweezers capable of manipulating many samples at once.

not interfere with the obstruction and reforms fully only a few micrometres further on. Garcés-Chávez *et al.* show that even when a Bessel beam of light is partially obstructed by an object in the first fluid-filled compartment, particles in a second, spatially separated, fluid compartment can be manipulated using the same light beam and with the same precision. This has been impossible with the conventional optical manipulation techniques used so far.

The authors created a diffraction-free Bessel beam⁵ using simple lenses and a special conical lens called an axicon. The bright spot at the beam's centre is a 'non-diffracting' focal line of light and is distinctly different from

'normal' optical trapping beams. The narrow beam has a propagation distance of a few millimetres, which is about 40 times the Rayleigh range (that is, the distance along which particles can be guided) of a normal, or 'gaussian', beam, and this means that optical manipulation can be done in multiple back-to-back compartments at the same time.

The Bessel beam can be imagined as a single rod of light, pushing the trapped particles along its axis. The trapping can extend over a few millimetres, and particles can be held firmly in two dimensions. In contrast, a gaussian beam can hold particles in a stable, three-dimensional configuration^{6,7}, but the beam diverges quickly, so particles can only be guided for a few micrometres along the axis of such a beam.

Certainly, one of the most impressive applications of (gaussian) optical tweezers is in the study of molecular motors^{8,9} and polymer mechanics¹⁰. These experiments have provided fresh insight into biochemical processes at the single-molecule level and are of great relevance in biology. The key to such experiments is in immobilizing the biological molecule of interest on the surface of a bead and attaching its interacting partner molecule to a second bead, which can then be trapped in three dimensions using infrared laser light⁶. But particles or beads caught in a Bessel beam are constantly pushed along the beam axis, and so comparable studies using Bessel-beam tweezers are not yet possible.

It is foreseeable, however, that if a counter-propagating Bessel beam were applied, the particles' forward motion could be stopped⁶. Like balancing a rod on the tip of another rod, a stable three-dimensional trap between the beams could be established if the power of the beams were adjusted properly. Extending this idea to exploit the Bessel-beam tweezers' ability to work on many particles simultaneously over considerable distance, patterns could be created from interfering Bessel beams, forming arrays of light spots that could be used, for example, to rotate microsystems, or to control lab-on-a-chip microstructures.

In biology, Bessel-beam tweezers could be used to sort the oval-shaped particles produced in the micro-dissection of chromatin (made up of DNA and proteins) through tiny openings or pores. For cell sorting, these tweezers are capable of more accurate guiding than systems used so far¹¹. The Bessel-beam technique will no doubt find application across a broad range of science. ■

Martin Hegner is at the National Centre of Competence in Research, Nanoscale Science, Institute of Physics, University of Basel, CH-4056 Basel, Switzerland.

e-mail: martin.hegner@unibas.ch

1. Garcés-Chávez, V., McGloin, D., Melville, H., Sibbett, W. & Dholakia, K. *Nature* **419**, 145–147 (2002).
2. Ashkin, A. *Phys. Rev. Lett.* **24**, 156–159 (1970).
3. Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E. & Chu, S.

- Opt. Lett.* **11**, 288–290 (1987).
 4. Ogura, Y., Kagawa, K. & Tanica, J. *Appl. Optics* **40**, 5430–5435 (2001).
 5. Durnin, J., Miceli, J. J. & Eberly, J. H. *Phys. Rev. Lett.* **58**, 1499–1501 (1987).
 6. Grange, W. *et al. Rev. Sci. Instrum.* **73**, 2308–2316 (2002).
 7. Lang, M. J., Asbury, C. L., Shaevitz, J. W. & Block, S. M. *Biophys. J.* **83**, 491–501 (2002).
 8. Kojima, H., Muto, E., Higuchi, H. & Yanagida, T. *Biophys. J.* **73**, 2012–2022 (1997).
 9. Smith, D. E. *et al. Nature* **413**, 748–752 (2001).
 10. Williams, M. C. & Rouzina, I. *Curr. Opin. Struct. Biol.* **12**, 330–336 (2002).
 11. Buican, T. N. *et al. Appl. Opt.* **26**, 5311–5316 (1987).

Cancer

Stuck at first base

Louise van der Weyden, Jos Jonkers and Allan Bradley

People with the genetic disease Peutz–Jeghers syndrome have many intestinal polyps — benign tissue outgrowths. These seldom become malignant, and the reason may lie in the properties of the affected gene.

Several inherited human diseases are characterized by the formation of polyps in the gut. Polyps are benign outgrowths of tissue with a disordered structure. In some of these diseases, the polyps undergo transformation into malignant gastrointestinal tumours — but this occurs relatively rarely in Peutz–Jeghers syndrome. In an effort to understand why, Bardeesy and colleagues¹ (page 162 of this issue) have produced mice carrying mutations in the mouse counterpart of the *LKB1* gene, which is affected in the human syndrome. The authors' studies of these animals have thrown up some unexpected findings.

Cancer, a disease characterized by unregulated cell proliferation, is caused by the stepwise accumulation of mutations in certain key genes. Every cancer is different, displaying a unique constellation of genetic changes. Yet one can identify some common mutational targets and molecular themes that lie at the heart of the transformation of a normal cell into a potentially deadly cancer cell. Thus, certain genes such as *RAS* and *p53* are often mutated in different cancers, suggesting that some deregulated cellular processes are common to many, if not all, tumours.

On the other hand, as our understanding of the molecular basis of tumour development increases, the richness of the mechanisms that have evolved to control cell growth is becoming more apparent. For example, studies of inherited mutations in tumour-suppressor genes (which usually keep the brakes on cell growth) show that some cell types are highly susceptible to malignant transformation, whereas others with the same mutations seem to be resistant. In addition, different cancer-predisposing mutations in the same cell type can have very different effects on malignancy.

Hereditary polyposis syndromes illustrate these phenomena² (Fig. 1). Intestinal polyps are generally composed of two cell types, epithelial and stromal cells; malignant tumours derived from such polyps consist mainly of transformed epithelial

cells. Patients with inherited mutations in the *APC* gene develop familial adenomatous polyposis, which is characterized by numerous epithelial-rich polyps that have a propensity to progress to malignant cancer. By contrast, patients with inherited inactivating mutations in *SMAD4* or *BMPRIA* develop juvenile polyposis, and people with *LKB1* mutations develop Peutz–Jeghers syndrome (PJS). Both syndromes are characterized by relatively benign, stromal-rich polyps called hamartomas. In juvenile polyposis, the inactivation of *SMAD4* is limited to the stromal cells, so the epithelial cells are influenced only indirectly³ (Fig. 1). But the limited malignant potential of polyps in PJS patients must have a different cause, because *LKB1* inactivation occurs in the epithelial, not the stromal, component⁴. Bardeesy *et al.*¹ now provide the first clues

to why *LKB1* mutations in the intestinal epithelium produce polyps that rarely become malignant.

To investigate the role of *LKB1* in human PJS, several groups^{5–7} have generated mice that lack the relevant mouse gene, *Lkb1*. In all cases, mice lacking both copies of the gene (homozygous animals) die as embryos. Heterozygous mice (with one deleted and one normal copy of the gene) survive, but develop gastrointestinal hamartomas with features akin to those in patients with PJS or juvenile polyposis. These animals are valuable mouse models of PJS. But the molecular basis for the limited malignant potential of PJS polyps has remained a conundrum, as the early death of the homozygous embryos prevents cell lines from being isolated and studied.

To tackle this issue, Bardeesy *et al.* generated mice that lacked one copy of *Lkb1* entirely and also carried a 'conditional' copy of the gene, which is inactivated only under specific experimental conditions (so the animals could develop normally). The authors isolated mouse embryonic fibroblast cells (MEFs) from these animals, and then switched off the conditional *Lkb1* gene in these cells, generating homozygous MEFs in culture.

MEFs are widely used for measuring two hallmarks of cancer: cell immortalization (the ability to multiply indefinitely in culture) and malignant transformation. The propensity of cultured normal MEFs to become immortal is limited by senescence — a programme of cellular 'ageing'. This crucial fail-safe mechanism helps to prevent transformation and is triggered by various

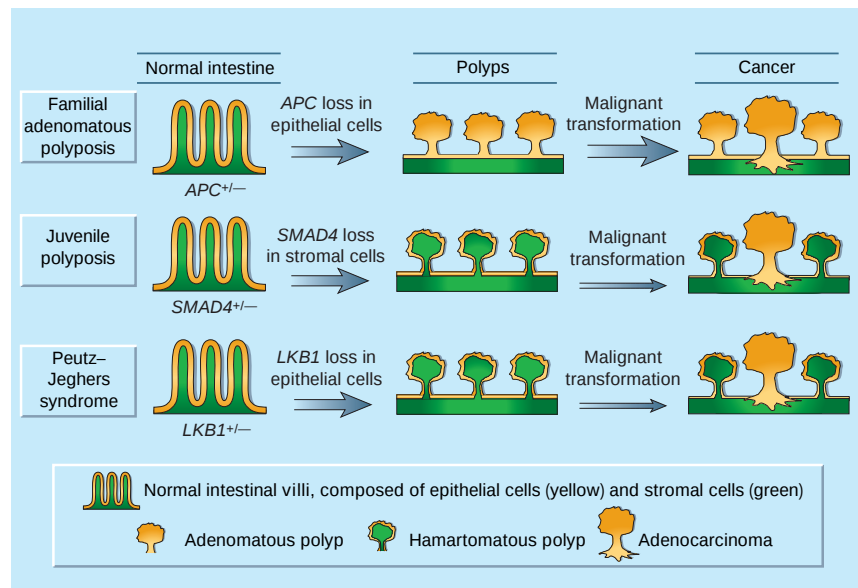


Figure 1 Development of intestinal cancer from intestinal polyps. Patients with the three diseases indicated on the left carry one normal and one mutant copy of the *APC*, *SMAD4* or *LKB1* tumour-suppressor genes (hence *APC*^{+/-}, and so on). Loss of the normal copy of the gene leads to the formation of polyps. Transformation of the epithelial cells in a polyp leads to development of adenocarcinomas, the relative frequency of which is indicated by the thickness of the arrow.

Type of mouse embryonic fibroblast cells	Early passage	Late passage	With activated RAS oncogene	With activated RAS and SV40 T antigen
Wild type	Growth	Senescence	Senescence	Transformation
No p53 or <i>Ink4a/ARF</i> tumour suppressors	Growth	Growth	Transformation	Transformation
No <i>Rb</i> and <i>p107</i> ; no <i>Rb</i> , <i>p107</i> and <i>p130</i>	Growth	Growth	Senescence	Transformation
No <i>Lkb1</i>	Growth	Growth	Senescence	No transformation

Figure 2 A different type of tumour suppressor? The features of the mouse *Lkb1* gene are rather unusual, as seen here. *Rb*, *p107* and *p130* are tumour-suppressor genes of the *Rb* family. SV40 is simian virus 40. Senescence prevents indefinite cell growth (immortalization); transformation refers to processes by which cells become cancer cells. 'Early' and 'late' passage refers to the relative amount of time cells have spent in culture.

stimuli, such as DNA damage, the expression of activated oncogenes (which promote cell growth), and stress induced by the tissue-culture process⁸.

Bardeesy *et al.* found that their homozygous MEFs were resistant to tissue-culture-induced senescence, and thereby became immortal, but still underwent senescence induced by DNA damage or oncogenes. This suggests that a lack of *Lkb1* renders cells immortal by inhibiting or alleviating the culture-stress-induced activation of senescence. Importantly, however, these immortal MEFs also turned out to be resistant to transformation induced by potent combinations of oncogenes, such as activated RAS and SV40 large T antigen, which readily transform normal cells (Fig. 2). So, although the lack of *Lkb1* causes immortalization, the immortalized cells seem to be resistant to subsequent transformation.

So *Lkb1* deficiency is a double-edged sword, somehow promoting perpetual cell growth but preventing malignant transformation — and explaining why the polyps in PJS patients develop yet remain benign. But how does *Lkb1* loss in the intestinal epithelium promote the development of stromal-rich polyps? Bardeesy *et al.* examined the genes expressed in the homozygous MEFs and in polyps from heterozygous mice, and detected marked increases in the expression of various genes encoding components of the extracellular matrix and secreted signalling molecules. So perhaps factors such as these, produced by *Lkb1*-deficient epithelial cells, influence the stromal cells in PJS polyps. In keeping with this idea, the authors observed that homozygous MEFs, and the medium in which these cells were cultured, affected the expression of specific genes in normal MEFs.

An unanswered question is why PJS patients are more likely than usual to develop other types of cancer. This seems odd, given that a lack of *LKB1* prevents malignant transformation of intestinal polyps. Does this mean that *LKB1*-deficient cells are resistant

to transformation by some activated oncogenes (such as members of the RAS pathway), but more susceptible to others? The low frequency of activating RAS mutations in tumours from PJS patients certainly hints

at this possibility. Alternatively, the order of events may matter. For example, *LKB1* inactivation in normal intestinal epithelium might protect those cells from malignant transformation, resulting in benign polyps. But in cell types that have already acquired mutations in other cancer genes, *LKB1* loss might promote tumour progression. No doubt Bardeesy and colleagues' mouse model will prove valuable in addressing these and other issues.

Louise van der Weyden, Jos Jonkers and Allan Bradley are at the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK.

e-mail: abradley@sanger.ac.uk

1. Bardeesy, N. *et al.* *Nature* **419**, 162–167 (2002).
2. Haggitt, R. C. & Reid, B. J. *Am. J. Surg. Pathol.* **10**, 871–887 (1986).
3. Kinzler, K. W. & Vogelstein, B. *Science* **280**, 1036–1037 (1998).
4. Wang, Z. J. *et al.* *J. Pathol.* **188**, 613–617 (1999).
5. Ylikorkala, A. *et al.* *Science* **293**, 1323–1326 (2001).
6. Miyoshi, H. *et al.* *Cancer Res.* **62**, 2261–2266 (2002).
7. Jishage, K. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 8903–8908 (2002).
8. Sherr, C. J. *Nature Rev. Mol. Cell Biol.* **2**, 731–737 (2001).

Ecology

Oceans under the microscope

Andrea Belgrano and James H. Brown

Phytoplankton are marine algae that support all ocean life, so it is important to understand the processes that control their distribution, abundance and diversity. Macroecology offers a way to do so.

Increasingly, scientists are turning to the relatively new approach of macroecology to study the structure and dynamics of complex ecosystems. Macroecologists look for statistical patterns in the abundance, distribution, biomass and diversity of individual organisms or species, in an effort to understand why these patterns emerge and what processes govern the structure and dynamics of the ecosystem as a whole. At first, macroecology relied heavily on large data sets of quite well-studied terrestrial birds and mammals^{1,2}. More recently, the approach and principles have been applied more widely — to fish, plants, molluscs and insects. Now, on page 154 of this issue, Li³ takes a macroecological approach to biological oceanography.

One problem in understanding the ecology of the oceans is the enormous spatial and temporal variation that results largely from climatic and physical oceanographic fluctuations (see, for example, Fig. 1). How can this problem be overcome so as to discover general features that are due to fundamental organizing processes? This is where the macroecological approach of looking at the statistical patterns of ecosystem components is so useful. For instance, to understand the structure and function of marine ecosystems, it is crucial to determine which

processes control the abundance, distribution and diversity of phytoplankton. These diverse, single-celled photosynthetic organisms, which convert solar energy into organic energy, are the primary producers that support the entire ocean food web — including commercially valuable fish stocks.

In his study, Li³ used flow cytometry to measure precisely the number and size of phytoplankton cells, as well as the amount of chlorophyll (the main photosynthetic pigment), in sea water. This technique made it practical for Li to quantify billions of cells, in thousands of samples from 23 oceanographic cruises, across a large area of the North Atlantic Ocean over a 13-year period.

Two especially important insights emerge from Li's data and analyses. First, he finds a strong, negative, 'allometric' scaling relationship — scaling that follows some power of body mass — between phytoplankton abundance (cell density, *N*) and cell volume (body mass, *M*). Specifically, $N \propto M^{-3/4}$. So, given that the rate of energy use by whole organisms (metabolism, *Q*) scales with body size as $Q \propto M^{3/4}$, the energy used by all of the phytoplankton cells, no matter which species, that make up a given size class must be the same as that used by all the phytoplankton that make up other size classes: $NQ \propto M^{-3/4}M^{3/4} \propto M^0$.

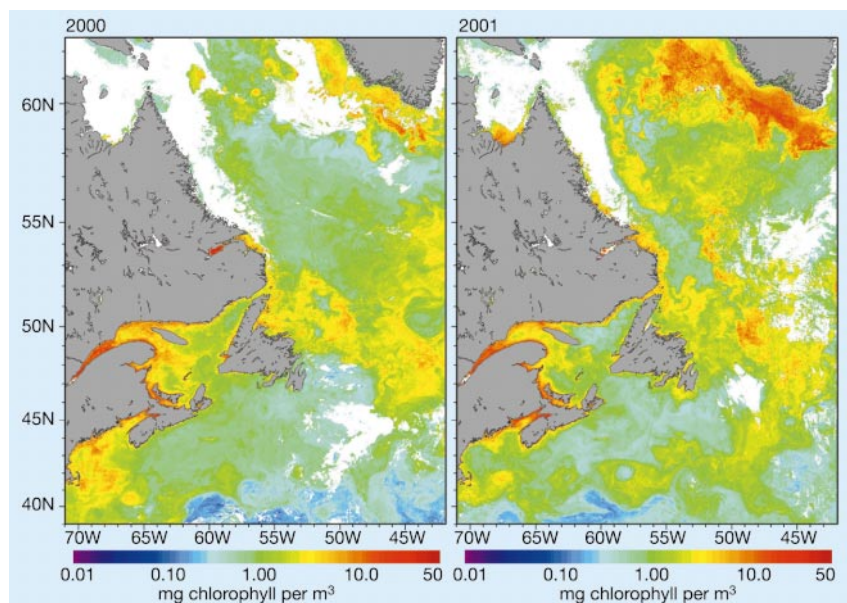


Figure 1 Biomass of chlorophyll from phytoplankton in surface waters of the northwest North Atlantic Ocean, the area investigated by Li³, in two-week periods in 2000 and 2001. The amount of chlorophyll is revealed by images produced from SeaWiFS data, collected by the OrbView-2 satellite, resolved to about 1.5 km per pixel. The images show broad similarities in chlorophyll concentration, as well as some significant differences, from year to year. They also show significant spatial heterogeneity. Li³ has used macroecological principles to try to disentangle such broad patterns, which probably result from large-scale climatic and physical oceanographic events, from the organizing processes inherent to marine phytoplankton. (These images are composites for 16 May to 30 May in 2000 and 2001, and are available at <http://oceanimage.mar.dfo-mpo.gc.ca>)

This extends to marine phytoplankton the 'energy equivalence rule' that has been demonstrated previously for other groups of organisms, including terrestrial animals and plants^{4,5}, suggesting that this rule is very general, if not universal. If all the individuals in a given size class do indeed use energy at the same rate as all the individuals that make up other size classes, this is an important feature of ecological organization that begs a mechanism. So far, it is not clear why organisms from different size classes should have access to, or be able to appropriate, equal quantities of energy.

Li's second important finding is that phytoplankton are most diverse at intermediate levels of chlorophyll concentration, and when the layers of the ocean water column are mixed to a degree somewhere between 'no mixing' and 'complete mixing'. Li obtained this result by measuring phytoplankton diversity not by tallying the number of species—as is traditional in ecology—but by creatively using flow cytometry to quantify the variety of phytoplankton types on the basis of their light-scattering properties. The results suggest that maximal diversity occurs when some intermediate degree of water mixing allows diverse functional types to coexist. (A high degree of mixing would allow one or a few species that can cope well with such disturbance to become dominant; likewise, no disturbance at all would foster competition, again allowing one or a few

superior competitors to become dominant.)

More generally, Li's work³ shows that, despite the enormous complexity and variety of marine phytoplankton, there are emergent

features of phytoplankton abundance and diversity that are related to cell size and overall chlorophyll content, respectively. The implication is that it is possible to make mechanistic connections between the metabolism of individual organisms (in this case, phytoplankton) and the roles of those organisms in ecosystems (here, in the productivity and feeding dynamics of oceans). For example, a recent study⁶ of marine food webs suggests how oceanographic and climatic events, through their effects on the abundance and productivity of phytoplankton, can cause fluctuations in commercial fish stocks.

Ecology has generally lacked unifying theories. Much of the emphasis has been on the seemingly large differences between different kinds of organisms and ecosystems, and on the extensive spatial and temporal variation within ecosystems. Now, however, there is increasing evidence that some macroecological patterns and mechanistic processes hold across diverse taxa and ecological systems. Such generality suggests exciting prospects for conceptual unification. ■

Andrea Belgrano and James H. Brown are in the Department of Biology, University of New Mexico, 167 Castetter Hall, Albuquerque, New Mexico 87131-1091, USA.
e-mail: belgrano@unm.edu

1. Brown, J. H. *Macroecology* (Chicago Univ. Press, 1995).
2. Gaston, K. J. & Blackburn, T. M. *Pattern and Process in Macroecology* (Blackwell, Oxford, 2000).
3. Li, W. K. W. *Nature* **419**, 154–157 (2002).
4. Damuth, J. *Nature* **290**, 699–700 (1981).
5. Enquist, B. J., Brown, J. H. & West G. B. *Nature* **395**, 163–165 (1999).
6. Stenseth, N. C. *et al. Science* **297**, 1292–1296 (2002).

Earth science

Baked Alaska

Peter Clift and Karen Bice

The warming of the Earth's climate more than 50 million years ago is as yet unexplained. Now the finger points to the heating of sediment in the Gulf of Alaska as an important source of the greenhouse gas methane.

Between about 58 and 52 million years ago, during the late Palaeocene and early Eocene epochs, the Earth's climate warmed considerably. What was the cause of this unusually long warming trend? Writing in *Geology*, Hudson and Magoon¹ put forward an idea that adds to thinking on the subject.

The marine geological record shows that Earth's climate has experienced swings of cooling and heating over various timescales. Change over thousands to hundreds of thousands of years seems to be related to orbital variations, but climate trends lasting millions of years have been harder to explain. The interval between the late Palaeocene and early Eocene has attracted particular interest, not only because of the exceptional

warming but also because this was a time of widespread and intense tectonic activity at the Earth's surface.

Attempts to find out whether the two are connected have usually centred on the release of greenhouse gases, especially CO₂, generated by tectonic activity. In some models the CO₂ produced is related to volcanism, especially in the northeast Atlantic². In others, the suggestion is that CO₂ was released from limestones originally deposited on the northern coast of the Indian subcontinent, which were buried, heated and then emitted CO₂ on a large scale when this landmass collided with the rest of Asia³.

Hudson and Magoon¹ propose an alternative tectonic trigger for the global warming of 58–52 million years ago, one that invokes

methane, not CO₂, as the key greenhouse gas. Their model deals with the Gulf of Alaska, where they suggest that large volumes of sediment, eroded from freshly uplifted mountains, were deposited in deep water off the 2,200-km coast of Alaska. In this region the oceanic Pacific plate is subducting below Alaska but the overlying sediments are largely scraped off against the edge of the continent to form a 'subduction-accretion complex' (Fig. 1). As sediment becomes buried and granite pushes up into the sedimentary layer, the heat leads to the production of hydrocarbons, including methane, from organic matter in the sediment. The large volumes of methane released to the atmosphere, increasing greenhouse-gas concentrations, might explain the higher global temperatures 58–52 million years ago.

Dating of events is crucial in assessing possible connections between tectonic events and greenhouse-gas emissions in the late Palaeocene and early Eocene. Take the case of the models invoking CO₂. Although volcanism in the northeast Atlantic began about 63 million years ago, the greatest volumes of lava did not erupt until about 56 million years ago⁴ — too late to trigger the initial warming. Analysis of the sedimentary rocks in the 'collision zone' between India and Asia⁵ shows that the earliest collision between these land masses happened only about 50 million years ago. In assessing Hudson and Magoon's hypothesis, a similar caveat might apply. Age constraints indicate that heating of the eastern part of Alaska⁶ might have occurred too late to have been important in generating methane.

Another aspect of the new model¹ deserving scrutiny is the estimated total methane production. Hudson and Magoon's estimate might be too high because it is based on a total sediment thickness of 25 km, assuming that 10 km has been removed by erosion. The eroded thickness has been estimated from radioisotope dating of the cooling of rocks through 300 °C, assuming a typical temperature gradient in the crust of 30 °C km⁻¹ (see, for example, ref. 6). However, such cooling could also be caused by the removal of the overlying rocks by faulting, not erosion. In this case, the 'missing' 10 km of sediment has not been lost but merely displaced sideways. In practice this would reduce the amount of organic material available to be cooked into methane. Nonetheless, so great is the amount of methane predicted by Hudson and Magoon¹ that even if production were halved this would still have resulted in a methane concentration in the atmosphere of several times the modern level.

Despite the apparent efficiency of Alaska as a generator of greenhouse methane, there is reason to wonder if this gas is really the likely climate culprit here. The popularity of methane as an agent in climate change has grown since its implication in the brief and

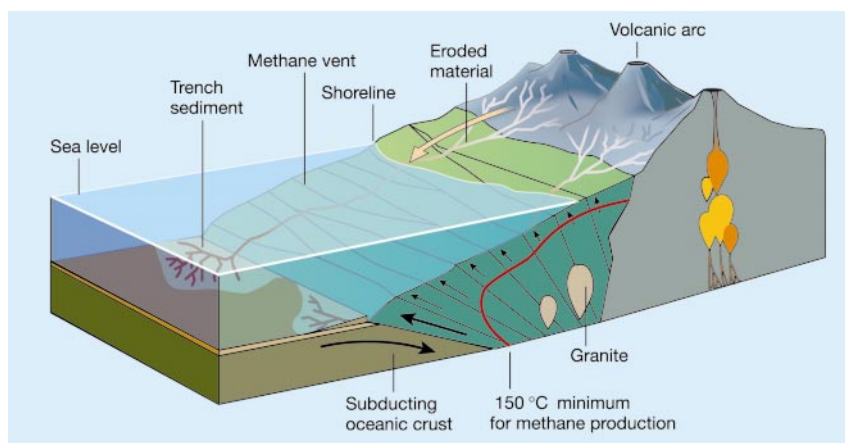


Figure 1 The Alaskan continental margin during the late Palaeocene epoch. Sediment washed down from newly lifted mountains collects at the sea floor. The subduction of the oceanic plate carries that sediment down into the Earth's crust, where, below the 150 °C level, the organic matter in it is converted into methane and other hydrocarbons. Hudson and Magoon¹ propose that the release of methane — a greenhouse gas — from this 'accretionary prism' of rock contributed to the global climate warming that happened over 50 million years ago.

dramatic Palaeocene–Eocene Thermal Maximum⁷ (around 55 million years ago), when temperatures in the deep oceans and at high latitude increased by 5–7 °C in less than 10,000 years. Here, the abrupt release of methane, previously stored in frozen form beneath the sea floor, provides a viable explanation for observed changes in marine and terrestrial carbon isotope records. But the direct climatic effects of methane that escapes to the atmosphere remain a matter of debate. Other new work⁸ shows that, because of its much longer residence time in the ocean–atmosphere system, CO₂ might be responsible for the formation of polar stratospheric clouds (whose insulating effects could explain the especially elevated temperatures at high latitudes^{1,9}). The uncertainty over this issue is all the more severe because most climate models currently lack the chemical detail and the resolution to predict the occurrence and radiative effects of these clouds.

There is also a long-standing problem in understanding the climate in continental interiors, for which models predict lower winter temperatures than those inferred from fossil evidence. Neither the direct radiative effects of increased greenhouse gases nor polar stratospheric clouds can account for this discrepancy. The cause of model underpredictions of temperatures both in polar regions and in continental interiors might lie in inadequate sensitivity of the modelled feedbacks from water vapour and cloud to increased greenhouse-gas concentrations. If so, then these same models are probably underestimating the strength of the hydrological cycle and the warming effects of the current increases in greenhouse-gas levels.

Hudson and Magoon's work¹ is notable in highlighting a possible overlooked influence on climate change. In attempting to understand past climate trends, episodes of faster

or slower carbon recycling at subduction–accretion complexes require more attention, as Hudson and Magoon have shown that in certain cases this process can influence the global greenhouse-gas budget to the same degree as does volcanic activity. Sedimentation and tectonic activity in the Gulf of Alaska is probably inadequate to explain the entire warming trend during the late Palaeocene and early Eocene by itself, although it might have contributed. Despite this advance in relating solid-Earth evolution to climate change, we are still struggling to understand the disparity between climate-model temperature estimates and those inferred from the fossil record, for polar regions and continental interiors during past warm intervals. A satisfactory, comprehensive mechanism has yet to be proposed. But when it emerges, we can hope that it will solve both model problems simultaneously, and apply more generally to warming episodes in the Earth's history.

Peter Clift and Karen Bice are in the Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.

e-mail: pclift@whoi.edu

- Hudson, T. L. & Magoon, L. B. *Geology* **30**, 547–550 (2002).
- Eldholm, O. & Thomas, E. *Earth Planet. Sci. Lett.* **117**, 319–329 (1993).
- Kerrick, D. M. & Caldeira, K. *Chem. Geol.* **108**, 201–230 (1993).
- Saunders, A. D., Fitton, J. G., Kerr, A. C., Norry, M. J. & Kent, R. W. in *Large Igneous Provinces* (eds Mahoney, J. J. & Coffin, M. L.) 45–94 (Am. Geophys. Union Monogr. 100, 1997).
- Clift, P. D., Carter, A., Krol, M. & Kirby, E. in *The Tectonic and Climatic Evolution of the Arabian Sea Region* (eds Clift, P. D., Kroon, D., Craig, J. & Gaedicke, C.) (Geol. Soc. Lond. Spec. Publ. 195, in the press).
- Sisson, V. B., Hollister, L. S. & Onstott, T. C. *J. Geophys. Res. B* **94**, 4392–4410 (1989).
- Dickens, G. R., O'Neill, J. R., Rea, D. K. & Owen, R. M. *Paleoceanography* **10**, 965–971 (1995).
- Kirk-Davidoff, D. B., Schrag, D. P. & Anderson, J. G. *Geophys. Res. Lett.* **29**, 10.1029/2002GL014659 (2002).
- Sloan, L. C. & Pollard, D. *Geophys. Res. Lett.* **25**, 3517–3520 (1998).

Mexican waves in an excitable medium

The stimulation of this concerted motion among expectant spectators is explained.

The Mexican wave, or *La Ola*, which rose to fame during the 1986 World Cup in Mexico, surges through the rows of spectators in a stadium as those in one section leap to their feet with their arms up, and then sit down again as the next section rises to repeat the motion. To interpret and quantify this collective human behaviour, we have used a variant of models that were originally developed to describe excitable media such as cardiac tissue. Modelling the reaction of the crowd to attempts to trigger the wave reveals how this phenomenon is stimulated, and may prove useful in controlling events that involve groups of excited people.

Using video recordings, we analysed 14 waves in football stadia holding over 50,000 people. The wave (Fig. 1) usually rolls in a clockwise direction and typically moves at a speed of about 12 metres (or 20 seats) per second and has a width of about 6–12 m (corresponding to an average width of 15 seats). It is generated by no more than a few dozen people standing up simultaneously, and subsequently expands through the entire crowd as it acquires a stable, near-linear shape. (For details and interactive simulations, see <http://angel.elte.hu/wave/>.)

Because of the relative simplicity of the Mexican wave, we were able to develop a quantitative treatment of this type of collective behaviour by building and simulating models that accurately reproduce and predict its details. We found that well-established approaches to the theoretical interpretation of excitable media^{1–3}, which



Figure 1 The Mexican wave, or *La Ola*, sweeping through a crowd of spectators. A few dozen fans leap up with their arms raised and then sit down as people in the next section jump to their feet to repeat and propagate the motion.

were originally created to describe processes such as forest fires or wave propagation in heart tissue, can be generalized to include human social behaviour.

We developed two mathematical simulation models, one minimal and one more detailed. By analogy with models of excitable media, people are regarded as excitable units — they can be activated by an external stimulus (a distance- and direction-weighted concentration of nearby active people exceeding a threshold value, c). Once activated, each unit follows the same

set of internal rules to pass through the active (standing and waving) and refractory (passive) phases before returning to its original resting (excitable) state.

The simpler model distinguishes only three states (excitable, active and passive) and accounts for variations in individual behaviour by considering the transition probabilities between these states. The more detailed model takes into account an actual deterministic activity pattern. The two versions differ in the way in which stochasticity (that is, differences and fluctuations) in the above behavioural patterns is represented (for details, see <http://angel.elte.hu/wave/>).

We applied these models to investigate the conditions that are necessary for triggering a wave. Figure 2a shows the evolution of a wave provoked by the simultaneous excitation (rising to their feet) of a small group of people. By using parameters deduced from video recordings for the different sizes and characteristic times of the phenomenon (interaction radius, reaction/activation times and probabilities), we were able to reproduce the size, form, velocity and stability of the wave.

Figure 2b shows the probability of generating a wave when a small group of varying size attempts to trigger it under different values of the excitation threshold. Our results indicate that the dependence of the eventual occurrence of a wave on the number of initiators is a sharply changing function — that is, triggering a Mexican wave requires a critical mass of initiators.

Our approach may have implications for

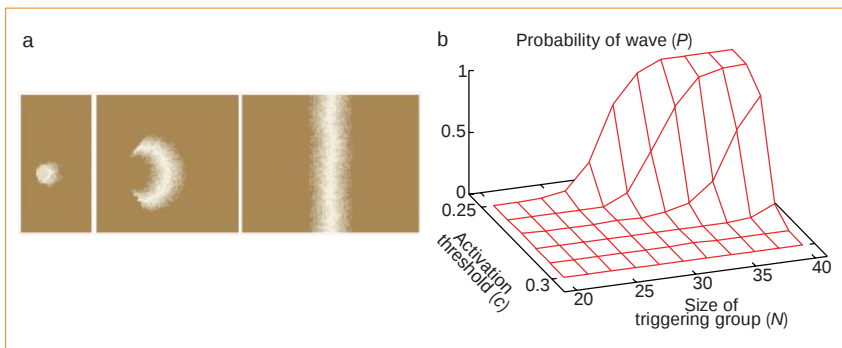


Figure 2 A model of the Mexican wave. If the weighted concentration of active people within a radius R around a person is above the person's threshold, c_i (chosen randomly from $[c - \Delta c, c + \Delta c]$), then that person becomes activated. Weights decrease exponentially with distance and change linearly with the cosine of the direction, so that people on the left of an individual have an influence that is w_0 times stronger than those on the right. The direction of the wave's motion is determined by this anisotropy because of the breakdown of spontaneous symmetry in the early stages as a result of anticipation and anisotropic perception (as most people are right-handed). **a**, Snapshots of the n -state model in which, after activation, a person deterministically goes through n_a active states (stages of standing up) and n_r refractory states. The wave is shown at 0.5 s, 2 s and 15 s after the triggering event, in a crowd with 80 rows of seats; brightness corresponds to the level of activity. Parameters are $n_a = n_r = 5$, $c = 0.25$, $\Delta c = 0.05$, $R = 3$ and $w_0 = 0.5$. **b**, The ratio of successful triggering events, $P(N, c)$, where P is the probability of a wave occurring as a function of the number of people, N , in the group who are trying to induce a wave, and the average threshold, c . Parameters are as in **a**; data represent means from 128 simulations.

crowd control. For example, in violent street incidents associated with demonstrations or sporting events, it is essential to understand the conditions under which small groups can gain control of the crowd, and how rapidly and in what form this perturbation or transition in behaviour could spread.

I. Farkas*, D. Helbing†, T. Vicsek*

*Department of Biological Physics, Eötvös

University Budapest, 1117 Budapest, Hungary

e-mail: vicsek@angel.elte.hu

†Institute for Economics and Traffic, Dresden

University of Technology, 01062 Dresden, Germany

1. Wiener, N. & Rosenbluth, A. *Arch. Inst. Cardiol. Mexico* **16**, 205–265 (1946).
2. Greenberg, J. M. & Hastings, S. P. *SIAM J. Appl. Math.* **34**, 515–523 (1978).
3. Bub, G., Shrier, A. & Glass, L. *Phys. Rev. Lett.* **88**, 058101 (2002).

Competing financial interests: declared none.

Metallurgy

High nickel release from 1- and 2-euro coins

The amount of nickel is regulated in European products that come into direct and prolonged contact with human skin¹ because this metal may cause contact allergy, particularly hand eczema^{2–4}. Here we show that 1- and 2-euro coins induce positive skin-test reactions in sensitized individuals and release 240–320-fold more nickel than is allowed under the European Union Nickel Directive. A factor contributing to this high release of nickel is corrosion due to the bimetallic structure of these coins, which generates a galvanic potential of 30–40 mV in human sweat.

We performed skin tests with 1- and 2-euro coins in seven patients known to have nickel-contact allergy. After 48 and 72 h with these coins fixed by transparent tape onto their skin, all seven patients showed a strong reaction, with erythema, infiltration and formation of vesicles; they showed no reaction to 1% zinc chloride in Vaseline or to 1% copper sulphate in water.

In a quantitative nickel-release test (the European Standard EN 1811; ref. 5), the 50-cent coin did not release a measurable amount of nickel, as expected. However, we found that the 1- and 2-euro coins released more nickel than pure nickel itself (Fig. 1).

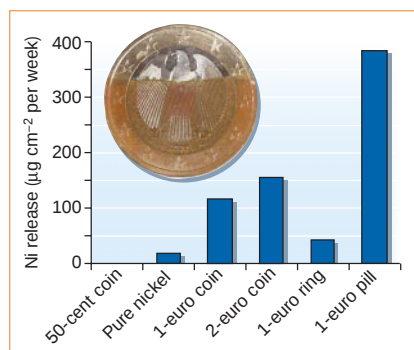


Figure 1 Release of nickel from euro coinage compared with that from pure nickel in artificial sweat, as measured by the EN 1811 standard reference test⁵ (values here have not been divided by 10, as specified for the test). Release of nickel from the bimetallic 1- and 2-euro coins is higher than from pure nickel. Inset, corrosion of a 1-euro piece after partial immersion in artificial sweat for 36 h.

This was particularly high from the inner component (the ‘pill’) of the 1-euro coin, but not for the outer component (the ‘ring’). These values are among the highest nickel-release rates ever measured on coins (see refs 6, 7 for a comparison).

In the 1-euro coin, the ring is made of a yellow alloy (‘nickel brass’) that consists of copper with 20% zinc and 5% nickel by weight; the white (‘cupro-nickel’) pill is copper with 25% nickel by weight; in the 2-euro coin the ring is cupro-nickel and the pill is nickel brass. As 1- and 2-euro coins are bimetallic, we measured the galvanic potential between the two metals with a high-impedance voltmeter after mechanically separating the pill and ring of a freshly minted coin and immersing them in either artificial sweat or saturated NaCl solution.

We found a difference in electrode potential between the two metals (the yellow metal was more negative and the white more positive) that was dependent on time and on the electrical resistance of the connector. After immersion for 24 h at ambient temperature in either solution, the potential difference between the two alloys stabilized at 40 mV (it was about 30 mV during the first 10 hours, then drifted slowly upwards) for a resistance of 100 kΩ.

It is well known that a current can enhance galvanic corrosion and thereby cause more nickel release. In thin irregular electrolyte layers such as sweat deposits, galvanic corrosion should occur primarily near the bimetallic junction because of the high resistance to lateral current flow in the thin layer. We measured electrochemically the relative rates of corrosion of each alloy, and found that the yellow metal dissolves at least five times faster than the white in the active-corrosion range (results not shown). Therefore, although the yellow component contains only one-fifth of the nickel of the white, its rate of nickel release is as high as that from the white component, or possibly higher, because the contact areas of the two alloys with the skin are about the same.

Corrosion of the 1-euro coin is visible after immersion for 36 hours in artificial human sweat: the colours changed to brown and the surface structure was damaged (Fig. 1, inset). No corrosion is evident, however, in a Swiss 1-franc coin, which consists of 25% nickel and 75% copper,

under these conditions (results not shown).

We conclude that the actual release of nickel from the present 1- and 2-euro coins exceeds the value acceptable for prolonged contact with human skin (as defined by European Union directive 94/27; ref. 1) by a factor of between 240 and 320 (Fig. 1). Whether or not this is acceptable by European standards hinges on the meaning of ‘prolonged’ contact. Further investigation is warranted not only into the epidemiological implications of such high-nickel-releasing coins but also into the factors that promote nickel release, such as the crevice between the pill and the ring — a potential corrosion site.

Frank O. Nestle*, Hannes Speidel†, Markus O. Speidel†

*Department of Dermatology, University of Zürich Hospital, Zürich 8901, Switzerland

e-mail: nestle@derm.unizh.ch

†Institute of Metallurgy, Department of Materials, ETH Zürich, Zürich 8092, Switzerland

1. European Parliament and Council Directive 94/27/EEC *Official Journal of the European Communities* (Brussels, 1994).
2. Andersen, K. E., Burrows, D. & White, I. R. in *Textbook of Contact Dermatitis* (eds Rycroft, R. J. G., Menné, T. & Frosch, P. J.) 418–420 (Springer, Berlin, 1995).
3. Gollhausen, R. & Ring, J. *J. Am. Acad. Dermatol.* **25**, 365–369 (1991).
4. Nielsen, N. H. et al. *Br. J. Dermatol.* **141**, 676–682 (1999).
5. European Standard EN 1811: 1998D *Official Journal of the European Communities* (Brussels, 1999).
6. Speidel, M. O. & Uggowitzer, P. J. in *Materials in Medicine* 191–208 (VDF Hochschulverlag AG, ETH Zürich, 1998).
7. Liden, C. & Carter, S. *Contact Dermatitis* **44**, 160–165 (2001).

Competing financial interests: declared none.

Gene therapy

Biological pacemaker created by gene transfer

The pacemaker cells of the heart initiate the heartbeat, sustain the circulation, and dictate the rate and rhythm of cardiac contraction¹. Circulatory collapse ensues when these specialized cells are damaged by disease, a situation that currently necessitates the implantation of an electronic pacemaker². Here we report the use of viral gene transfer to convert quiescent heart-muscle cells into pacemaker cells, and the successful generation of spontaneous, rhythmic electrical activity in the ventricle *in vivo*. Our results indicate that genetically engineered pacemakers could be developed as a possible alternative to implantable electronic devices.

In the early embryonic heart, each cell possesses intrinsic pacemaker activity. The mechanism of spontaneous beating in the early embryo is remarkably simple³ — the opening of L-type calcium channels causes depolarization, and the subsequent opening of transient outward potassium channels leads to repolarization. As development

progresses, the heart differentiates into specialized functional regions, each with its own distinctive electrical signature. The atria and ventricles become electrically quiescent, and the overall rate and rhythm is determined by a small number of pacemaker cells situated within compact 'nodes'.

We investigated the possibility that pacemaker activity is latent in adult ventricular myocytes and is normally repressed by the inward-rectifier potassium current (I_{K1}). This current, which is intensely expressed in the atrium and ventricle but not in nodal pacemaker cells, is encoded by the *Kir2* gene family⁴. I_{K1} stabilizes a strongly negative resting potential and should therefore suppress excitability. We tested this idea by producing dominant-negative inhibition of *Kir2*-encoded inward-rectifier potassium channels in the ventricle.

Replacement of three amino-acid residues in the pore structure of *Kir2.1* (introduction of alanine residues at positions 144–146: GYG144–146AAA, here designated as *Kir2.1AAA*) creates a dominant-negative construct that suppresses current when expressed together with wild-type *Kir2.1* (refs 5, 6). We packaged *Kir2.1AAA* and green fluorescent protein (GFP) into a bicistronic adenoviral vector (AdEGI–*Kir2.1AAA*) and injected this construct into the left ventricular cavity of guinea-pigs (*Cavia porcellus*) during transient cross-clamping of the great vessels; transduction was successful in about 20% of ventricular myocytes, as determined by GFP fluorescence. Myocytes isolated 3–4 days after *in vivo* transduction with *Kir2.1AAA* showed roughly 80% suppression of I_{K1} .

Non-transduced left ventricular myocytes isolated from AdEGI–*Kir2.1AAA*-injected animals, as well as GFP-positive cells from control AdEGI-injected hearts, showed no spontaneous activity, but did generate single action potentials when subjected to depolarizing external stimuli (Fig. 1a). By contrast, *Kir2.1AAA*-transduced myocytes exhibited one of two phenotypes: a stable resting potential from which prolonged action potentials could be elicited by external stimuli (7 of 22 cells; results not shown), or spontaneous activity (Fig. 1b).

The spontaneous activity, which was observed in all cells in which I_{K1} was suppressed below 0.4 pA/pF (at -50 mV), resembles that of genuine pacemaker cells: the maximum diastolic potential (-60.7 ± 2.1 mV, 15 out of 22 *Kir2.1AAA* cells, $P < 0.05$, *t*-test) is relatively depolarized, with repetitive, regular electrical activity initiated by gradual 'phase-4' depolarization and a slow upstroke^{1,7}. *Kir2.1AAA*-transduced pacemaker cells responded to β -adrenergic stimulation (by isoprenaline) just as nodal cells do, by increasing their pacing rate⁷.

Electrocardiography revealed two pheno-

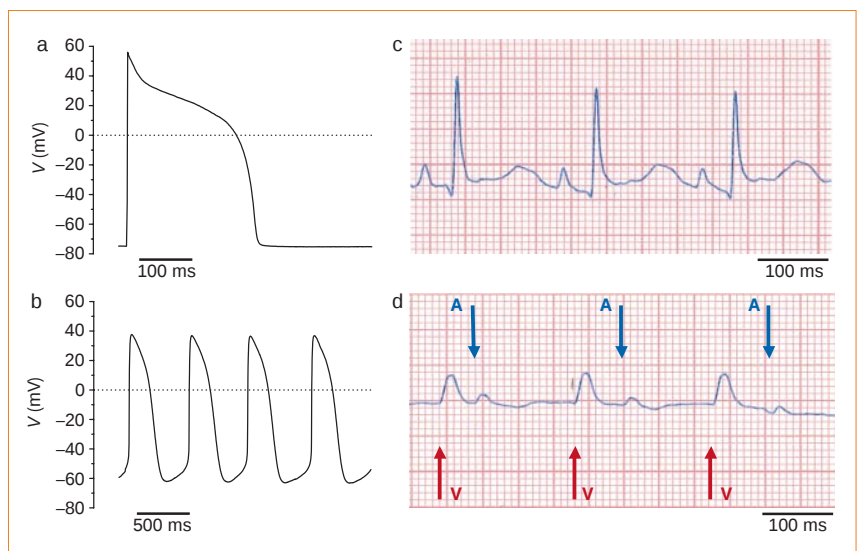


Figure 1 Suppression of *Kir2.1* channels unleashes pacemaker activity. **a**, Stable action potentials evoked by depolarizing external stimuli in control ventricular myocytes. **b**, Spontaneous action potentials in *Kir2.1AAA*-transduced myocytes with suppressed inward-rectifier potassium current (I_{K1} , pacemaker phenotype). **c**, Baseline electrocardiograms in normal sinus rhythm (control). **d**, Ventricular rhythms for the pacemaker phenotype 72 h after transduction of *Kir2.1AAA*. P waves (A, arrow) and wide QRS complexes (V, arrow) 'march through' to their own rhythm.

types *in vivo*: normal sinus rhythm (Fig. 1c) with simple prolongation of the QT interval (results not shown), or an altered cardiac rhythm indicative of spontaneous ventricular foci. In 40% of animals after transduction with *Kir2.1AAA*, premature beats of ventricular origin could be distinguished by their broad amplitude, and were found to 'march through' to a beat that was independent of, and more rapid than, that of the physiological sinus pacemaker (Fig. 1d). In these proof-of-concept experiments, the punctate transduction required for pacing occurred by chance rather than by design, in that the distribution of the transgene throughout the ventricles was not controlled. Nevertheless, the foci of induced pacemakers caused the heartbeat to originate from the ventricles.

Our findings provide new insight into the biological basis of pacemaker activity. Such activity has been thought to require the highly localized expression in nodal cells of 'pacemaker genes', such as those of the *HCN* family⁸, although an important role for I_{K1} has also been recognized⁷. Exposure to barium induces automaticity in ventricular muscle and myocytes because of its time- and voltage-dependent blocking of I_{K1} (refs 9, 10). However, barium also permeates L-type calcium channels in mixed solutions of Ca^{2+} and Ba^{2+} (ref. 11) and slows inactivation¹², effects that make it difficult to interpret barium's effect strictly in terms of I_{K1} . Our dominant-negative approach is durable and regionally specific; the barium effect is not.

Our direct experimental evidence indicates that specific suppression of *Kir2* channels can give rise to pacemaker activity in

ventricular myocytes. The crucial factor for pacing is the absence of the strongly polarizing I_{K1} , rather than the expression of certain genes (although such genes may play an important modulatory role in genuine pacemaker cells)¹³. As genetic suppression of I_{K1} converts myocytes to pacemaker cells, localized delivery of constructs such as *Kir2.1AAA* to the myocardium may eventually be useful in the creation of biological pacemakers for therapeutic purposes.

Junichiro Miale*, Eduardo Marbán, H. Bradley Nuss*

Institute of Molecular Cardiology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

e-mail: marban@jhmi.edu

**Present address: Cardiology Research Group, Department of Medicine/Division of Cardiology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA*

- Brooks, C. M. & Lu, H.-H. *The Sinoatrial Pacemaker of the Heart* (Thomas, Springfield, Illinois, 1972).
- Kusumoto, F. M. & Goldschlager, N. N. *Engl. J. Med.* **334**, 89–97 (1996).
- Wobus, A. M., Rohwedel, J., Maltsev, V. & Hescheler, J. *Ann. NY Acad. Sci.* **752**, 460–469 (1995).
- Kubo, Y., Baldwin, T. J., Jan, Y. N. & Jan, L. Y. *Nature* **362**, 127–133 (1993).
- Herskowitz, I. *Nature* **329**, 219–222 (1987).
- Slesinger, P. A. *et al. Neuron* **16**, 321–331 (1996).
- Irisawa, H., Brown, H. F. & Giles, W. *Physiol. Rev.* **73**, 197–227 (1993).
- Santoro, B. & Tibbs, G. R. *Ann. NY Acad. Sci.* **868**, 741–764 (1999).
- Imoto, Y., Ehara, T. & Matsuura, H. *Am. J. Physiol.* **252**, 325–333 (1987).
- Hirano, Y. & Hiraoka, M. *J. Physiol. (Lond.)* **395**, 455–472 (1988).
- Rodríguez-Contreras, A., Nonner, W. & Yamoah, E. N. *J. Physiol. (Lond.)* **538**, 729–745 (2002).
- Campbell, D. L., Giles, W. R. & Shibata, E. F. *J. Physiol. (Lond.)* **403**, 239–266 (1988).
- Brown, H. F., Kimura, J., Noble, D., Noble, S. J. & Taupignon, A. *Proc. R. Soc. Lond. B* **222**, 329–347 (1984).

Competing financial interests: declared none.

Radionuclide contamination

Nanometre-size products of uranium bioreduction

One strategy that is being pursued to tackle the international problem of actinide contamination of soils, sediments and water is to use microbial activity to 'fix' these radionuclides into an insoluble form that cannot be readily dispersed. Here we show that uraninite (UO_2) particles formed from uranium in sediments by bacterial reduction are typically less than 2 nanometres across and that the small size has important implications for uraninite reactivity and fate. Because these tiny particles may still be transported in an aqueous environment, precipitation of uranium as insoluble uraninite cannot be presumed to immobilize it.

The discovery that bacteria can convert mobile hexavalent uranium, U(VI) , to tetravalent uranium, U(IV) , which precipitates as uraninite¹, has spurred investigations into the *in situ* bioremediation of uranium-contaminated groundwater. These strategies assume that the formation of highly insoluble uraninite² will inhibit the mobilization of uranium.

We used organic substrates to create anaerobic conditions to stimulate growth of native bacteria in complex natural uranium-contaminated sediments and water, taken from the Midnite Mine, Washington, USA. After 1 month, the aqueous uranium concentration decreased from 20 to 0.3 parts per million. Synchrotron-based uranium L-III edge X-ray absorption near-edge spectroscopy (XANES) analysis of the incubated sediment indicated that most U(VI) had been reduced to U(IV) . Selected-area electron-diffraction patterns and crystallographic analysis of two-dimensional transmission electron microscope (TEM) lattice-fringe images confirmed the structure as being that of uraninite (which is similar to that of calcium fluoride).

We also constructed a gene library encoding small-subunit ribosomal RNAs (16S rRNAs) from the incubated sediment³ to identify the bacterium that is likely to be responsible for the conversion of uranium to uraninite. Out of 94 clones, 42 clustered strongly with *Desulfosporosinus* spp. (results not shown), a Gram-positive sulphate-reducing microbe. We confirmed that *Desulfosporosinus* spp. isolated from the sediment was able to reduce U(VI) to U(IV) and to precipitate U(IV) as uranium oxide (manuscript submitted).

Uraninite formed by mixed cultures in the incubated natural sediment, as well as by pure cultures of *Desulfosporosinus* spp., was characterized by high-resolution TEM (Fig. 1a). The uraninite crystals are less than 3 nm in diameter and occur as discrete

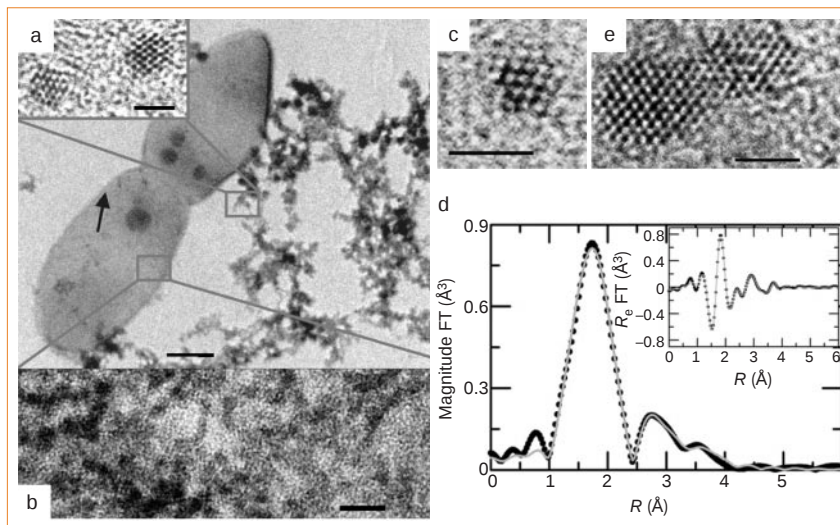


Figure 1 Characterization of bioreduced uraninite (UO_2) nanoparticles. **a**, Transmission electron microscopy (TEM) image of flocculated UO_2 nanoparticles associated with *Desulfosporosinus* spp. bacteria (arrow). Inset, high-resolution TEM image of isolated particles. **b**, TEM image of the *Desulfosporosinus* cell surface coated with UO_2 nanoparticles (size, 1.5–2.5 nm). **c**, A single uraninite particle (about 1.3 nm across). **d**, Magnitude of the Fourier transform (FT) of X-ray-absorption fine-structure (XAFS) data obtained from the sediment, and the best-fit model. R , radial distance (uncorrected for electron phase shift). Inset, real part of the Fourier transform. R_e , real part of Fourier-transformed data. **e**, Growth of uraninite crystals formed by orientated attachment of individual nanoparticles. Scale bars: 0.6 μm (**a**, inset, 2 nm), 6 nm (**b**) and 2 nm (**c**, **e**).

(Fig. 1a, inset) and aggregated particles, and as particle coatings on bacterial cell surfaces (Fig. 1b). Many crystalline particles were less than 1.5 nm in diameter (Fig. 1c), and our images indicated that even smaller particles were present (results not shown).

Extended X-ray-absorption fine-structure (EXAFS) analysis (Fig. 1d) of the incubated sediment revealed that the average uraninite U–U coordination number was 5.6 ± 4 (compared with 12 in bulk uraninite). This confirms that a large proportion of uranium ions in nanoparticles are at the surfaces of the uraninite nanoparticles, as expected for tiny particles. Structure models consistent with the EXAFS-determined coordination number imply an average particle size of about 1.5 nm, which is close to the size of the smallest crystals detected in our high-resolution TEM images. We conclude that bacterial U(VI) reduction produces uraninite particles in a range of sizes that extends down to molecular-scale clusters.

EXAFS data indicate that all U(IV) in the sediment was 8-coordinated by oxygen, as in the bulk uraninite structure. U–U distances (0.380 nm) are 0.007 nm shorter in biogenic uraninite than in bulk uraninite, implying that there may be significant displacement of surface U(IV) ions. The 0.007-nm contraction in U–U spacing corresponds to a surface stress of about 4.82 N m^{-1} , which represents more than a billion-fold increase in the solubility product compared with well-crystallized uraninite⁴ (based on the reaction $\text{UO}_2 + 2\text{H}_2\text{O} = \text{U}^{4+} + 4\text{OH}^-$; the actual value depends on the quantitative relationship between surface stress and surface energy).

The nanometre-scale size of biogenic

uraninite particles is likely to influence both its reactivity and transport. We would expect these UO_2 crystals to be mobile in porous sediments and to re-oxidize rapidly. Flocculation, on the other hand, should reduce the transport potential. TEM images indicate that crystals grow from orientated aggregates of uraninite (Fig. 1e), which may be mediated in some cases by extracellular organic polymers (our unpublished results). To predict the likely environmental fate of UO_2 nanoparticles, it will first be necessary to determine the transport physics of isolated nanoparticles and of nanoparticles adsorbed onto colloidal particles and organics, and to define models that link crystallization with crystal-growth kinetics, flocculation kinetics and fluid flow.

Yohey Suzuki*§, Shelly D. Kelly†, Kenneth M. Kemner†, Jillian F. Banfield*‡

*Department of Geology and Geophysics, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA

e-mail: jill@seismo.berkeley.edu

†Environmental Research Division, Argonne National Laboratory, Argonne, Illinois 60439, USA

‡Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA

§Present address: Japan Marine Science and Technology Centre, Yokosuka 237-006, Japan

- Lovely, D. R., Phillips, E. J. P., Gorby, Y. A. & Landa, E. R. *Nature* **350**, 413–416 (1991).
- Grethe, I. et al. *Chemical Thermodynamics of Uranium* (Nuclear Energy Agency, North-Holland, The Netherlands, 1992).
- Bond, P. L., Smriga, S. P. & Banfield, J. F. *Appl. Environ. Microbiol.* **66**, 3842–3849 (2000).
- Rai, D., Felmy, A. R. & Ryan, J. L. *Inorg. Chem.* **29**, 260–264 (1990).

Competing financial interests: declared none.

***RAD6*-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO**

Carsten Hoege*, Boris Pfander*, George-Lucian Moldovan*, George Pyrowolakis*† & Stefan Jentsch*

* Department of Molecular Cell Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18a, 82152 Martinsried, Germany

The *RAD6* pathway is central to post-replicative DNA repair in eukaryotic cells; however, the machinery and its regulation remain poorly understood. Two principal elements of this pathway are the ubiquitin-conjugating enzymes *RAD6* and the *MMS2*–*UBC13* heterodimer, which are recruited to chromatin by the RING-finger proteins *RAD18* and *RAD5*, respectively. Here we show that *UBC9*, a small ubiquitin-related modifier (SUMO)-conjugating enzyme, is also affiliated with this pathway and that proliferating cell nuclear antigen (PCNA)—a DNA-polymerase sliding clamp involved in DNA synthesis and repair—is a substrate. PCNA is mono-ubiquitinated through *RAD6* and *RAD18*, modified by lysine-63-linked multi-ubiquitination—which additionally requires *MMS2*, *UBC13* and *RAD5*—and is conjugated to SUMO by *UBC9*. All three modifications affect the same lysine residue of PCNA, suggesting that they label PCNA for alternative functions. We demonstrate that these modifications differentially affect resistance to DNA damage, and that damage-induced PCNA ubiquitination is elementary for DNA repair and occurs at the same conserved residue in yeast and humans.

DNA is highly vulnerable to spontaneous and environmental damage in living cells. If not repaired, DNA damage may cause mutation-induced genetic variation, ageing, carcinogenesis, or cell death. Eukaryotic DNA repair pathways are highly conserved from yeast to humans. Three principal pathways are classified according to genetic relations of yeast (*Saccharomyces cerevisiae*) DNA-repair mutants (for reviews see refs 1–3). The *RAD3* group mediates nucleotide excision repair, the *RAD52* group directs double-strand break repair through homologous recombination, whereas the *RAD6* group functions in post-replication repair. Proteins of the *RAD6* group seem to act on the stalled replication machinery that has encountered a damaged template, thereby accomplishing repair and allowing replication to resume^{1,4}. This repair can be achieved by at least two different *RAD6*-dependent mechanisms: one mechanism is considered error-prone as it uses specialized translesion polymerases that insert correct or incorrect nucleotides across a damaged site; the other *RAD6*-dependent mode is error-free because it uses the information of the undamaged sister duplex at the replication fork.

The mechanism by which the *RAD6* pathway controls the different modes of post-replication DNA repair has remained enigmatic. Previous work has shown that modification of proteins by conjugation to ubiquitin is pivotal for *RAD6*-dependent DNA repair. *RAD6* itself encodes a ubiquitin-conjugating enzyme, which attaches ubiquitin to substrate proteins in collaboration with ubiquitin-activating enzyme⁵. Furthermore, two other members of the *RAD6* group, *UBC13* and *MMS2*, form a heterodimeric ubiquitin-conjugating enzyme, which catalyses the formation of non-canonical multi-ubiquitin chains linked via K63 of ubiquitin^{6–11}. Both enzymes, *RAD6* and *UBC13*–*MMS2*, are recruited to chromatin by interaction with the RING-finger-containing, DNA-binding proteins *RAD18* and *RAD5*, respectively^{9,12}. We have demonstrated previously that the two distinct ubiquitin-conjugating enzymes associate by means of a *RAD18*–*RAD5* interaction, and we proposed that this assembly might modify proteins by K63-linked multi-ubiquitin chains⁹. In contrast to K48-linked multi-ubiquitination, mono-ubiquitination and

modification by K63-linked chains does not generally promote proteasomal degradation, but rather it seems to alter the function of the substrate or to mediate protein–protein interactions¹³. Nevertheless, the key question of how ubiquitination affects DNA repair has not been answered so far, because the proteins relevant to DNA repair that are ubiquitinated by the *RAD6* pathway have remained unidentified.

Modification of PCNA by SUMO

During a study of the protein modification system that uses the small ubiquitin-related modifier, SUMO, we obtained unexpected insights into the *RAD6* pathway. This type of modification is widespread and affects a large number of cytosolic and nuclear proteins (for reviews see refs 14–16). SUMO, a protein that shares about 18% sequence identity with ubiquitin, has been suggested to mediate protein–protein interaction or to function as a ubiquitin antagonist.

To investigate this modification in yeast we searched for new SUMO substrates. We immuno-purified SUMO–protein conjugates by a two-step protocol and identified the proteins by mass spectrometric analysis of tryptic peptides (Fig. 1a). Peptide masses of one conjugate with a relative molecular mass of 45,000 (*M*_r, 45K) matched those of tryptic fragments predicted for the gene product of *POL30*; that is, the yeast proliferating cell nuclear antigen PCNA. Further analysis verified that a fraction of the protein is modified by SUMO *in vivo* (Fig. 1b). PCNA, a protein of 29K, is loaded onto DNA strands as a trimeric ring where it functions as a clamp and processivity factor for DNA polymerases^{17,18}. It is essential for a variety of S-phase functions including replication, replication-linked repair and silencing. These activities are matched by PCNA levels, which rise at S phase¹⁹. We synchronized cells using α -factor and observed that SUMO conjugation to PCNA is regulated by the cell cycle and that it precedes the rise of PCNA during S phase (Fig. 1c). This finding suggests that SUMO modification of PCNA is important for events that take place during S phase.

Modification sites

Similarly to ubiquitin, SUMO is conjugated to ϵ -amino groups of

† Present address: Biocenter, University of Basel, 4056 Basel, Switzerland.

lysine residues of the target protein^{14,15}. To identify the acceptor residue(s) for SUMO conjugation, we changed all 18 lysine residues of yeast PCNA individually to arginine. Only mutants at lysine 127 (K127R) and 164 (K164R) specifically altered the pattern of SUMO conjugates in a total yeast lysate (data not shown). Pull-down experiments using wild-type and mutant PCNA verified that yeast PCNA can be modified by SUMO conjugation at both residues and that a K127R K164R PCNA double mutant is not modified by SUMO to detectable levels (Fig. 1d). K127 lies within a postulated SUMO-modification consensus site (ψ KxD/E, where ψ is a hydrophobic residue; ref. 20), but appears to be unique for yeast PCNA. In contrast, K164 is a 'non-consensus' SUMO-conjugation site, yet this site is highly conserved in PCNA from yeast to humans. The

conserved K164 residue seems to be the prominent site for SUMO conjugation, whereas SUMO conjugation at K127 is stimulated when K164 has been experimentally mutated (Fig. 1d). Both modifications can occur on the same molecule, as doubly modified PCNA species can be detected in western blots of PCNA and SUMO (data not shown).

PCNA is also involved in DNA damage response and has been linked to the RAD6 pathway of post-replicative DNA repair²¹. When we treated cells with 0.02% of the DNA-damaging agent methyl methanesulphonate (MMS), we observed little change in the level of PCNA-SUMO conjugates (Fig. 2a; band S¹⁶⁴). However, when we applied a lethal dose of MMS (0.3%), there was marked induction of SUMO conjugation of PCNA, in particular at residue K164 (Fig. 2a, b). Thus, SUMO conjugation of PCNA is not triggered exclusively by S-phase events, but also by severe DNA damage.

Modification of PCNA by ubiquitin

We also noticed that treatment of cells with 0.02% MMS resulted in the induction of additional PCNA-specific bands, which were also present in *ubc9-1* mutants defective for SUMO conjugation (Fig. 2a–c; bands U1 and U2). From this finding and the sizes of the proteins we speculated that these species might be PCNA-ubiquitin conjugates. Indeed, as verified by pull-down experiments, we found that PCNA is modified by ubiquitination and that species carrying one, two, three and more ubiquitin moieties can be detected (Fig. 2b–d). Notably, mono- and multi-ubiquitination of PCNA does not seem to occur to detectable levels in untreated cells, or during S phase (Figs 1d and 2a–d), but only after treatment with a sublethal dose of DNA-damaging agents (0.02% MMS; Fig. 2a–d). Ubiquitination of PCNA is absent after treatment with 0.3% MMS for several hours, a condition that induces extensive SUMO conjugation at K164 (Fig. 2a, b). By using PCNA mutants that lacked specific lysine residues we noticed that ubiquitin is attached to residue K164 (Fig. 2b, c). Thus, ubiquitination and the principal type of SUMO conjugation target the same conserved lysine residue of PCNA, yet the two distinct modifications are regulated differentially.

PCNA is a substrate of the RAD6 pathway

In view of the results indicating that PCNA is ubiquitinated on DNA damage, we speculated that enzymes of the RAD6 pathway are involved. Indeed, when we monitored PCNA by western blot analysis we noticed that the ubiquitin modifications of PCNA were completely absent in *rad6* deletion mutants, whereas SUMO modification of PCNA was not affected (Fig. 2c). The same specific defect was observed when we used a mutant that lacked RAD18, a DNA-binding protein that recruits RAD6 to chromatin¹². Previously, we showed that RAD18 can associate with RAD5, which in turn recruits the heterodimeric UBC13–MMS2 ubiquitin-conjugating enzyme to chromatin, and we proposed that RAD6 and RAD18 might collaborate with UBC13–MMS2 and RAD5 in ubiquitination reactions⁹. In fact, ubiquitination of PCNA was affected equally in *ubc13*, *mms2* and *rad5* mutants (Fig. 2c). Only the multi-ubiquitinated PCNA species (two or more ubiquitin moieties) were absent in these mutants, whereas the mono-ubiquitinated PCNA species (band U1) was still present. As the UBC13–MMS2 complex catalyses multi-ubiquitin chains with K63 linkages⁷, we also looked for PCNA ubiquitination in a strain²² that expresses ubiquitin solely as a variant that lacks K63 (*ubiK63R*). Indeed, multi-ubiquitination of PCNA was also absent in this strain, whereas mono-ubiquitination still occurred (Fig. 2c, d). Together, these findings are consistent with a model in which DNA damage induces RAD6- and RAD18-dependent mono-ubiquitination of PCNA at K164, and that UBC13–MMS2 and RAD5 are required to attach additional ubiquitin moieties to the conjugate, thereby forming a K63-linked multi-ubiquitin chain (see Supplementary Information, Fig. 1).

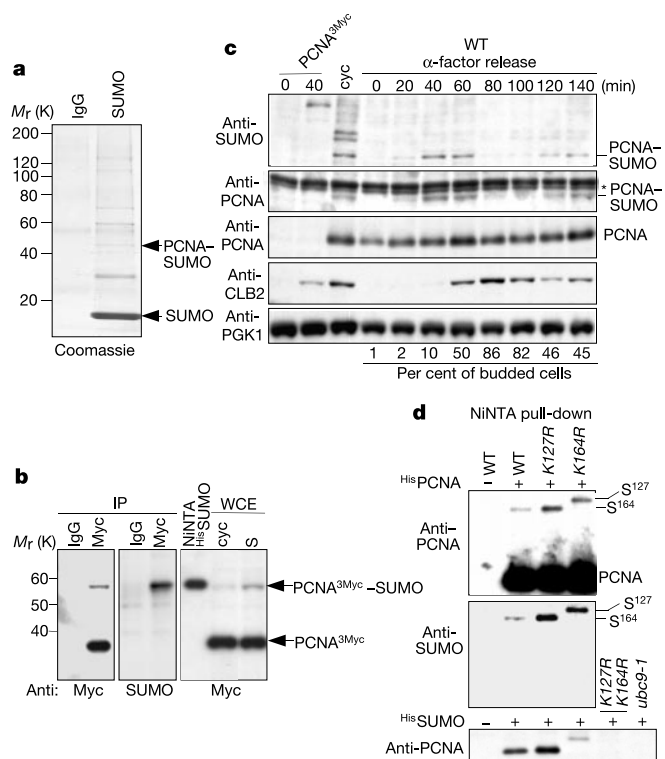


Figure 1 PCNA is modified by SUMO in *S. cerevisiae*. **a**, Two-step purification of His-tagged SUMO (HisSUMO) conjugates involving NiNTA pull-down followed by anti-SUMO affinity chromatography. The 45K PCNA-SUMO conjugate was identified by mass spectrometry. **b**, Anti-PCNA (tagged with a triple Myc epitope, PCNA^{3Myc}) immunoprecipitations from a *ulp1ts* strain defective in SUMO deconjugation (IP), NiNTA pull-downs from HisSUMO-expressing cells under denaturing conditions (NiNTA HisSUMO), and whole-cell extracts (WCE) of cycling wild-type (WT) cells (cyc) or S-phase cells (S). Proteins were detected by western blotting using antibodies as indicated below. **c**, PCNA SUMO modification is upregulated in S phase. Cultures were synchronized by α -factor arrest/release, and analysed by western blotting using the indicated antibodies. The identity of PCNA and PCNA-SUMO bands is demonstrated by the corresponding up-shift when Myc-tagged PCNA (PCNA^{3Myc}) was expressed as the sole source of PCNA (left two lanes). Relevant parts of the same gel are shown. Levels of PCNA-SUMO peak at 40 min, whereas those of unmodified PCNA peak at 60 min after release from α -factor arrest. Nonspecific bands are indicated by an asterisk. Cell-cycle stages were monitored by CLB2 cyclin levels, budding index (numbers at bottom), and FACS analysis (not shown). PGK1 served as a loading control. **d**, SUMO modification sites of PCNA. NiNTA pull-down (as above) from cells expressing HisPCNA (WT), His-tagged PCNA mutants (K127R and K164R) as the sole source of PCNA, or HisSUMO. Proteins were identified by western blotting as indicated. SUMO is preferentially attached to K164 (S¹⁶⁴), whereas modification at K127 (S¹²⁷) is induced when K164 is absent (right lane). PCNA is not modified by SUMO when both lysines are absent (K127R K164R) or when a *ubc9-1* strain was used (bottom panel).

Physical interactions

PCNA is loaded onto DNA by a clamp loader and requires homo-trimerization. We noticed that a PCNA mutant protein *pol30-52* (ref. 23) that is partially defective in trimerization was much less modified *in vivo* (data not shown). One possible interpretation of this finding is that the modifications take place after trimerization and that DNA-associated enzymes are involved. Indeed, when we tested all enzymatic components of the RAD6 ubiquitination system in yeast two-hybrid assays we discovered that only the DNA-binding proteins RAD18 and RAD5 interact with PCNA (Fig. 2e). Importantly, both proteins contain a RING finger, a hallmark of a class of ubiquitin ligases²⁴. This observation, together with our finding that RAD18 and RAD5 mediate PCNA ubiquitination, indicates that the two proteins function as DNA-bound ubiquitin ligases for PCNA.

The SUMO-conjugating enzyme UBC9 usually binds substrates directly^{16,25}. This is also apparently the case for PCNA, as two-hybrid assays revealed a UBC9–PCNA interaction (Fig. 2f). Of note, UBC9 additionally binds RAD18 and RAD5 (Fig. 2e), and we confirmed these interactions by co-immunoprecipitations (Fig. 2g). This indicates that UBC9 is affiliated with the RAD6 pathway and that enzymes for ubiquitin and SUMO conjugation may communicate in a regulatory unit. Importantly, the observed physical association, together with our finding that ubiquitination and SUMO conjugation targets the same lysine residue, suggests that the distinct modifications label PCNA for specific functions.

Functions

To investigate the significance of our finding that PCNA is a

substrate of the RAD6 DNA repair pathway, we determined PCNA mutant phenotypes. Yeast cells that express, from the original gene locus, PCNA variants lacking either the ubiquitination/SUMO-conjugation site K164 (*pol30-K164R*) or the SUMO-conjugation site K127 (*pol30-K127R*) exhibit no discernible growth defects, indicating that the modifications are not essential for replication (Fig. 3a). However, we observed that *pol30-K164R* cells were highly sensitive to ultraviolet light and MMS. This phenotype was specific for *pol30-K164R* cells, as all of the other 17 lysine mutants of PCNA were not hypersensitive (data not shown). Interestingly, when K164 and K127 were absent simultaneously (*pol30-K164R K127R*), cells were less sensitive than the *pol30-K164R* strain. Thus, we conclude that modification at K164 is crucial for DNA repair, whereas SUMO conjugation at K127 seems to inhibit the DNA repair process.

Notably, deletion mutants of RAD6, RAD18 and RAD5 are more sensitive to ultraviolet light than the *pol30-K164R* mutant (Fig. 3c; see also Supplementary Information, Fig. 2), indicating that the encoded proteins have additional substrates or functions. This is not surprising as RAD6 has several cellular roles²⁶ and RAD5 is related to SNF2/SWI2 helicases²⁷. We found that *pol30-K164R* is epistatic to *mms2* (that is, *mms2* deletion does not confer additional sensitivity to *pol30-K164R*), indicating that K63-linked multi-ubiquitination of PCNA is firmly linked to MMS2-dependent DNA repair (Fig. 3c; see also Supplementary Information, Fig. 2). As *ubc13* and *mms2* mutants are specifically defective in error-free repair^{6–9}, this finding also demonstrates that K63-linked multi-ubiquitination of PCNA is elementary for this branch of RAD6-dependent repair. Moreover, the finding that an *mms2 pol30-K164R* double mutant is more

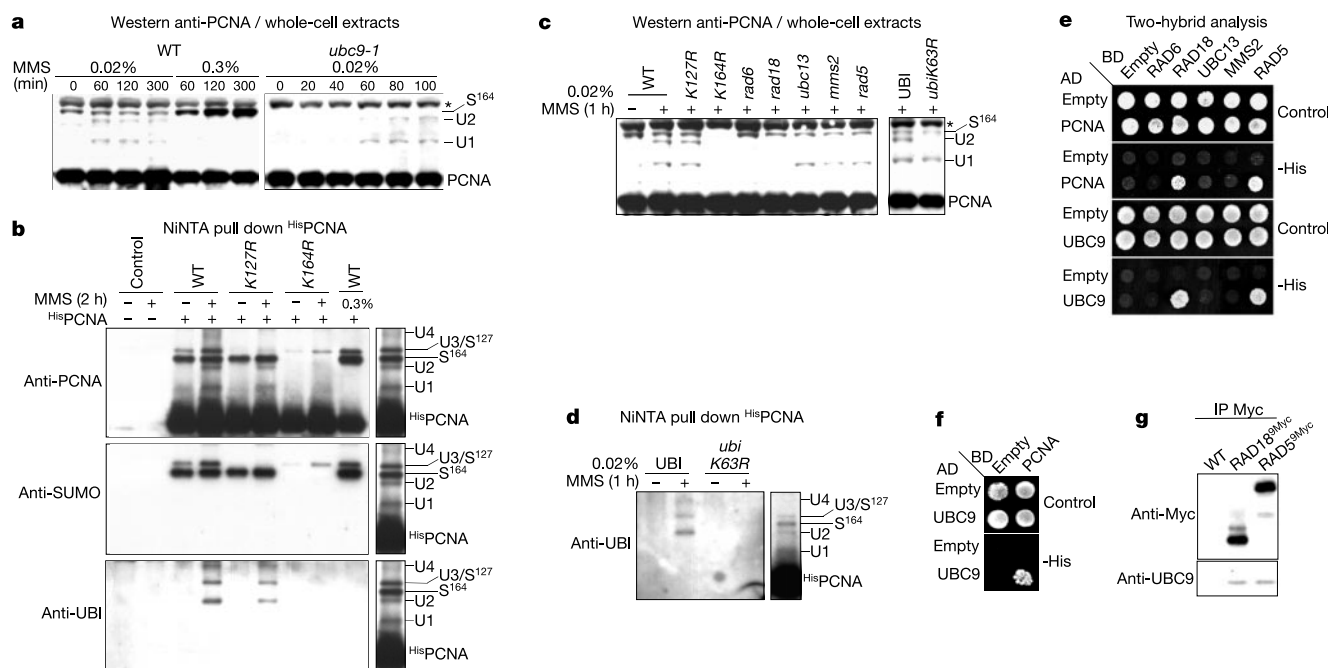


Figure 2 Regulation of PCNA SUMO modification and ubiquitination by DNA damage and link to the RAD6 pathway. **a**, PCNA modifications in WT and *ubc9-1* cells after treatment with MMS (0.02% and 0.3%) at 30 °C. Anti-PCNA western blots identify SUMO-modified (S) and ubiquitinated (U) species. Asterisk, nonspecific band co-migrating with forms U3 and S¹²⁷ (see below). **b**, Identification of SUMO-modified and ubiquitinated forms of PCNA after HisPCNA pull-down followed by western blotting. The mono-ubiquitinated form U1 is not recognized by the monoclonal antibody (but see **c** and Supplementary Information, Fig. 1). Strains and conditions are as in Fig. 1d; band assignment is given on the right (U1, mono-ubiquitinated; U2–4, multi-ubiquitinated; S¹²⁷ and S¹⁶⁴, SUMO-conjugated at K127 and K164, respectively). PCNA mutants (*K127R* and *K164R*) are used to show the lysines involved in conjugation. Forms U3 and S¹²⁷ migrate together in gels. Ubiquitination

is specifically induced by MMS. **c**, PCNA ubiquitination depends on the RAD6 pathway. Ubiquitination is absent in *rad6* and *rad18* mutants. Only mono-ubiquitinated PCNA (U1) is detected in *ubc13*, *mms2* and *rad5* (left panel), as well as in cells that express mutant ubiquitin (*ubiK63R*; right panel). SUMO modification is not affected by these RAD6 pathway mutants. **d**, Multi-ubiquitin chains on PCNA are linked via K63 of ubiquitin. Pull-downs of HisPCNA from WT (UBI) and *ubiK63R* mutant cells followed by anti-ubiquitin western blotting. **e**, **f**, Two-hybrid interactions between UBC9, PCNA and proteins from the RAD6 pathway. Fusion with either activating domain (AD) or DNA-binding domain (BD) of GAL4 are indicated. **g**, Co-immunoprecipitation of UBC9 (*Y1lac211::GAL-UBC9*) with chromosomally expressed Myc-tagged RAD18 (*RAD18^{9myc}*) or RAD5 (*RAD5^{9myc}*) detected by anti-UBC9 western blotting.

sensitive than the *mms2* single mutant suggests that mono-ubiquitination of PCNA at K164 (which can still occur in *mms2* single mutants) supports DNA damage tolerance, although less efficiently than multi-ubiquitination. We also investigated the effect of an *srs2* mutation on *pol30-K164R* mutant cells. *SRS2* encodes a DNA helicase, and loss of function of this enzyme partially suppresses the ultraviolet sensitivities of mutants defective in *RAD6*-dependent, error-free repair, by channelling repair to the *RAD52* pathway^{28–31}. We found a detectable suppression of the sensitivity of *pol30-K164R* mutants by the *srs2* mutation (Fig. 3c), confirming that K63-linked multi-ubiquitination of PCNA is linked to the error-free branch of *RAD6*-dependent DNA repair.

Additional double mutant analysis showed that the *pol30-K164R* mutation strongly enhanced the ultraviolet sensitivities of *rad2* and *rad52* mutants (Fig. 3c), suggesting that the PCNA modifications are relevant neither for *RAD2*-dependent excision repair nor for *RAD52*-dependent double-strand break repair. Interestingly, the ultraviolet sensitivities of *rad6*, *rad18* and *rad5* mutants are partially suppressed by the *pol30-K164R* mutation. As wild-type PCNA, but not PCNA-K164R, can be modified by SUMO at K164 in *rad6*, *rad18* and *rad5* single mutants (Fig. 2c), this finding also points towards a DNA repair inhibitory effect of PCNA modification by SUMO at K164 (in addition to K127; see above) in cells deficient in these enzymes.

Importantly, overexpression of wild-type PCNA renders *mms2* and *rad5* mutants—as well as cells that express a ubiquitin variant that is unable to form K63-linked chains (*ubiK63R*)—less sensitive to MMS (Fig. 3d). We conclude from this finding that K63-linked UBC13–MMS2 and *RAD5*-dependent multi-ubiquitination of PCNA activates the role of PCNA in DNA repair, and that high levels of possibly mono-ubiquitinated PCNA can compensate for a defect in PCNA multi-ubiquitination. Notably, overexpression of *pol30-K164R* does not suppress the sensitivity of these cells but rather seems to be detrimental for repair in these mutant and wild-type cells, indicating that non-ubiquitinated PCNA does not significantly support DNA repair. Indeed, the MMS sensitivities of *rad6* and *rad18* mutants are only barely suppressed by overexpression of wild-type PCNA (which can still be SUMO-conjugated at K164) but are suppressed slightly better by mutant *pol30-K164R*. These findings confirm that ubiquitination of PCNA is elementary for post-replicative DNA repair and indicate that SUMO modification of PCNA at K164 not only prevents PCNA ubiquitination but also inhibits PCNA-dependent DNA repair.

Human PCNA ubiquitination is linked to DNA damage

PCNA is a highly conserved protein with residues sharing 35% identity between yeast and human orthologues. Because of this similarity we investigated whether human PCNA is also modified.

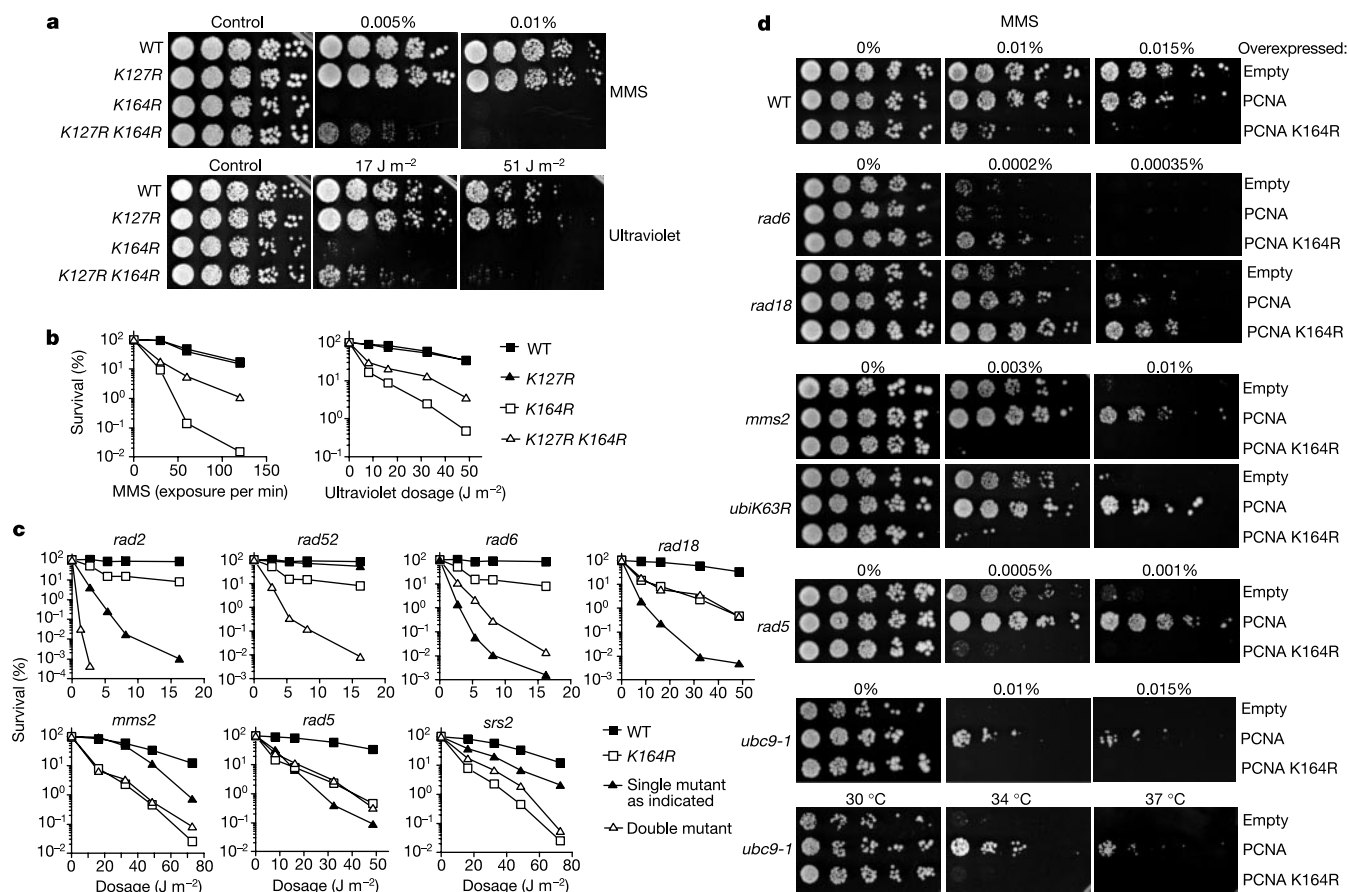


Figure 3 Role of PCNA modifications in DNA repair. **a**, Sensitivity of cells chromosomally expressing WT and mutant PCNA to MMS and ultraviolet light. Fivefold serial dilutions of an equal number of stationary-phase cells were spotted onto plates. DNA damage was induced by either MMS or ultraviolet light. PCNA mutants lacking K127 or K164, or both lysines, were used. **b**, Quantification of survival rates (as above) determined by plating. Values are presented as the percentage of surviving cells after treatment. **c**, Epistatic analysis of the PCNA mutant lacking residue K164 (*pol30-K164R*) with mutants of DNA

repair pathways. Quantification of survival rates after ultraviolet irradiation (as in **b**). Values are from two to four independent experiments with duplicated counts. The genotype to which *pol30-K164R* was crossed is indicated above the panels. **d**, PCNA overexpression from a high-copy plasmid (WT copy was retained in the genome) differentially suppresses MMS sensitivity of *RAD6* pathway mutants. Overexpression of WT PCNA and mutant PCNA (K164R) was compared to control (empty vector). Cells were spotted as above.

We could not detect SUMO modification or ubiquitination of human PCNA in HeLa cells under normal growth conditions. However, when we applied 0.02% MMS to the culture medium, we observed the appearance of an additional PCNA-specific band (Fig. 4a). This protein was verified by western analysis and pull-down experiments to be mono-ubiquitinated PCNA. Notably, the modification site K164 of the *S. cerevisiae* protein is conserved and found at identical positions in PCNA from yeasts, plants and higher eukaryotes, including humans (Fig. 4b). We replaced K164 of human PCNA by an arginine residue and expressed His-tagged forms of mutant and wild-type PCNA in HeLa cells by transfection. We observed that wild-type PCNA, but not mutant PCNA-K164R, was modified by ubiquitination in the presence of MMS. Thus, ubiquitination of PCNA at site K164 is conserved from yeast to mammals and is linked to the DNA damage response.

Discussion

PCNA is positioned at the crossroads of several replication-linked pathways. It is involved in leading and lagging strand DNA synthesis, cell cycle arrest, replication-linked DNA silencing, mismatch repair, nucleotide- and base-excision repair, and post-replicative error-free and error-prone DNA repair^{3,17,18,32}. PCNA directly associates with various DNA polymerases and functions as a sliding clamp, thereby stimulating accurate and processive DNA synthesis. In addition, PCNA seems to function as a platform for accessory factors of replication-linked functions. Previous studies have shown that PCNA is controlled by at least two distinct mechanisms. One important principle seems to be regulated loading/unloading of the trimeric PCNA ring on the DNA template by clamp loader (replication factor C). A second mechanism may involve PCNA inhibition through interaction with repressing molecules, the most prominent being the cell cycle checkpoint protein p21 (refs 17, 18).

We have shown that PCNA is also exquisitely modulated by covalent post-translational modifications involving ubiquitin and the ubiquitin-related protein SUMO. PCNA is a target for SUMO modification, mono-ubiquitination, and K63-linked multi-ubiquitination. We identified two target sites for modification: K164, a

conserved modification site that is both modified by SUMO and ubiquitin, and K127, a yeast-specific site that seems to be exclusively modified by SUMO. Both residues are positioned distally from the encircled DNA on the outside rim of the trimeric PCNA ring³³ (Fig. 5). K164 is located in one of the protruding tips of PCNA, whereas K127 lies within a large loop that connects the two domains of a PCNA monomer. This connecting loop mediates interaction with polymerases (refs 34, 35), suggesting that SUMO conjugation at this site might interfere with polymerase binding.

SUMO modification of PCNA is activated during S phase and modifies predominantly the conserved residue K164 and to a lesser extent K127 of yeast PCNA. Importantly, our data indicate that SUMO conjugation inhibits the role of PCNA in repair and suggests a function in conjunction with normal DNA replication. Modification of PCNA by SUMO is also markedly induced by a lethal concentration of MMS (Fig. 2a), but whether this upregulation is linked to DNA repair or to polymerase stalling is currently unclear. We also noticed that modification of PCNA by SUMO is mediated by SIZ1, a SUMO ligase of the SIZ/PIAS family³⁶ (see Supplementary Information, Fig. 3). This finding demonstrates that modification of PCNA by SUMO does not support DNA repair because *siz1* mutants are not sensitive to DNA-damaging agents³⁶. A possible clue with respect to the function of SUMO-modified PCNA comes from our finding that not only the MMS sensitivity of the *ubc9-1* mutant, but also its growth defect at high temperatures can be suppressed partially by overexpression of wild-type PCNA, but not the K164R mutant (Fig. 3d). This finding emphasizes the crucial importance of PCNA for the SUMO pathway even in the absence of DNA-damaging agents, again pointing to an important role of SUMO-modified PCNA in normal S phase.

Ubiquitination of PCNA is clearly linked to DNA damage and strictly depends on the *RAD6* pathway of DNA repair. It specifically targets the conserved K164 residue of yeast and human PCNA. Moreover, we show that PCNA ubiquitination is elementary for *RAD6*-dependent DNA repair. Previous data have indicated a complex orchestration of *RAD6*-dependent functions and revealed

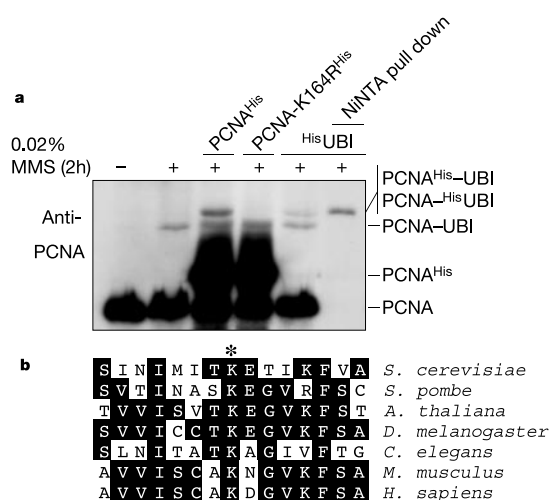


Figure 4 Human PCNA is ubiquitinated at the conserved K164 residue on DNA damage. **a**, Western blot against PCNA from HeLa cells. Cells were either untransfected (two left lanes) or transfected with constructs expressing His-tagged versions of human WT PCNA (PCNA^{His}) or K164R mutant PCNA (PCNA-K164R^{His}). Cultures were treated with 0.02% MMS for 2 h as indicated. Anti-PCNA antibodies detect endogenous (PCNA) and introduced His-tagged PCNA (PCNA^{His}). For the last two lanes, a construct expressing His-tagged UBI was transfected, and samples were loaded on a gel before and after NINTA pull-down. **b**, K164 is conserved in PCNA from yeast to humans. Shown is an alignment between residues 157 and 171 of yeast PCNA. K164 is marked by an asterisk.

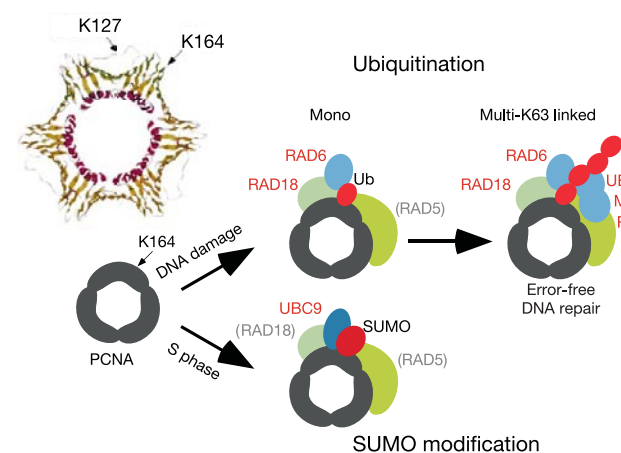


Figure 5 Model for ubiquitination and SUMO modification of PCNA. The structure of yeast PCNA is derived from ref. 33 (top left). K127 and K164 modification sites are indicated. DNA damage induces mono-ubiquitination of PCNA at K164, which is catalysed by ubiquitin-activating enzyme UBA1 (not shown), RAD6 and RAD18. DNA damage induces nuclear import of UBC13 and MMS2, which are recruited to chromatin by RAD5. The enzyme assembly catalyses K63-linked multi-ubiquitination of PCNA. In the absence of DNA-damaging agents, PCNA is modified by SUMO during S phase, involving SUMO-activating enzymes UBA2/AOS1 (not shown), UBC9, and the SUMO ligase SiZ1 (not shown). UBC9 interacts directly with PCNA and, perhaps transiently, also with RAD18 and RAD5. Enzymes required for the modification reactions are labelled in red; associated proteins are grey and in parentheses. Ubiquitin- and SUMO-conjugating enzymes are in blue; ubiquitin ligases are in green.

distinct pathways for error-free and error-prone modes of repair⁴. These pathways may be subdivided into distinct error-free branches and an error-prone mode that uses the translesion polymerases REV1/Pol ζ (encoded by *REV3* and *REV7*). Another RAD6/RAD18-dependent translesion polymerase is Pol η (encoded by *RAD30*)^{8,37–39}. Our findings provide a new conceptual framework for the function of the RAD6 pathway as it suggests that the stalled replication machinery may be switched to different modes of repair through distinct PCNA modifications.

We provide strong evidence that multi-ubiquitination of PCNA at K164 is pivotal for the error-free branch of RAD6-dependent DNA repair. We also show that two ubiquitin-conjugating enzymes, RAD6 and UBC13–MMS2, and two RING-finger ubiquitin ligases, RAD18 and RAD5, are involved in PCNA ubiquitination. Importantly, our findings suggest a mechanism by which these enzymes collaborate in a sequential manner (Fig. 5). Triggered by DNA damage, RAD6 appears to initially mono-ubiquitinate PCNA at residue K164. This reaction additionally requires RAD18, which recruits RAD6 to DNA-bound PCNA. DNA damage also induces nuclear translocation of UBC13 and MMS2, which results in an association of the heterodimeric ubiquitin-conjugating enzyme with chromatin-bound RAD5 (ref. 9). Moreover, through RAD5–RAD18 interaction, UBC13–MMS2 is brought into contact with RAD6. In a second enzymatic reaction, this assembly of two ubiquitin-conjugating enzymes and two ubiquitin ligases (together with ubiquitin-activating enzyme) seems to catalyse the conjugation of additional ubiquitin molecules onto the ubiquitin moiety of mono-ubiquitinated PCNA, thereby forming K63-linked multi-ubiquitin chains. These chains, unlike K48-linked chains, do not promote proteasomal degradation, and, indeed, we found no evidence for accelerated turnover of ubiquitinated PCNA (data not shown). In contrast, our data indicate that K63-linked multi-ubiquitination activates PCNA for being engaged in error-free repair. Post-replicative, error-free repair is thought to involve a transient template switch to the undamaged sister chromatid³. An attractive speculation is, therefore, that the ubiquitin chains on PCNA may stimulate interaction of the stalled replication machinery with proteins associated with the undamaged sister duplex. Interestingly, recent studies have indicated that K63-linked multi-ubiquitination involving UBC13–MMS2 can also activate certain protein kinases^{40,41}. Moreover, K63-linked multi-ubiquitination has a function in ribosome activity⁴² and endocytosis⁴³. It will be important to identify the parallels between these roles in order to understand the mechanism through which K63-linked multi-ubiquitin chains function.

Mono-ubiquitination also seems to be relevant for DNA repair (Fig. 3c). Because error-free repair requires K63-linked multi-ubiquitination^{7,9,44}, mono-ubiquitination of PCNA is expected to mediate another, possibly error-prone, mode of repair. Perhaps mono-ubiquitinated PCNA binds specific repair factors, such as polymerases involved in translesion synthesis. Mono-ubiquitination of PCNA seems to be prevalent in human HeLa cells, yet multi-ubiquitinated forms can be detected at lower levels (G.-L.M. and S.J., unpublished data). Importantly, mammals have homologues of UBC9, SIZ1, RAD6, RAD18, UBC13, MMS2 and possibly RAD5. Given this conservation, it seems probable that the mechanisms and functions are highly conserved, and, indeed, genetic studies have confirmed that human RAD18 and MMS2 are crucial for DNA repair^{45,46}.

This study also revealed an affiliation of the SUMO-modification system with the ubiquitin-conjugating machinery. We have demonstrated that SUMO modification and ubiquitination of PCNA are physically and functionally linked. Importantly, ubiquitination and SUMO modification target the same K164 residue of PCNA, demonstrating that the two modifications are mutually exclusive. SUMO-conjugating enzyme UBC9 not only interacts with PCNA but also with the ubiquitin ligases RAD18 and RAD5; however, this

interaction is not required for SUMO modification of PCNA (Fig. 2c), but it might be crucial for regulation. The most plausible model is that the ubiquitin/SUMO-conjugating enzymes are part of a regulatory switchboard, which directs PCNA for alternative functions. Several previous investigations have provided evidence for a link between SUMO modification and ubiquitination¹⁵. In cases where these two processes target the same lysine residue, studies suggest that SUMO might function as an antagonist for K48-linked multi-ubiquitination, thereby inhibiting proteasomal degradation of the substrate⁴⁷. Our data suggest that SUMO modification can also antagonize the non-proteolytic role of ubiquitin. Importantly, SUMO modification of PCNA at K164 does not merely block ubiquitination at this site, but it seems to direct the substrate for other functions. As both SUMO modification and ubiquitination are reversible processes^{14,15}, a sequential cycle of SUMO modification/ubiquitination for substrates can be envisioned. It seems attractive to speculate that the ubiquitin/SUMO switch discovered here might be a prevalent regulatory mechanism. □

Methods

Yeast techniques and cloning

The strains that we used (grown at 30 °C, if not stated otherwise) are isogenic to strains described in ref. 9, and have the genotypes as described in the text and figures. Yeast His-tagged SUMO was expressed from the *ADH* promoter and the integrative vector *YIplac211*. His-tagged UBI was expressed from a 2 μ plasmid with *CUP1* promoter by the addition of 100 μ M CuSO₄. Yeast PCNA (*POL30*) was cloned into the 2 μ -based plasmid *YEplac195* for overexpression studies or the *YCplac22* centromeric plasmid to create point mutations. Seven His codons were placed between the promoter and open reading frame (ORF) of *POL30* by a polymerase chain reaction (PCR)-based technique to create His-tagged PCNA in *YCplac22*. The shuffle strain *pol30::kanMX4 YCplac33* PCNA was used to create strains that express variants of PCNA or His-tagged PCNA as the only source of PCNA by selection on 5-fluoro-orotic acid. For Fig. 2d we expressed His-tagged PCNA in wild-type (UBI) or *ubiK63R* strain background. PCNA lysine mutants were integrated by homologous recombination into the *POL30* locus. Chromosomally tagged strains and mutants, if not already described^{9,48}, were constructed by a PCR-based strategy⁴⁹. The *ubc9-1* mutant strain (Y0233) contains an additional *bar1* knockout. For two-hybrid analysis, complete ORFs of *POL30* and *UBC9* were cloned into pGAD vectors. Assay and other RAD6 pathway clones were described previously⁹. Details of strains and plasmids will be provided on request.

Protein techniques and antibodies

Protein conjugates with His-tagged SUMO from 6 litres of cells with an absorbance at 600 nm of 1 (*A*_{600nm}) of the *ulpts* strain⁵⁰ (grown at 23 °C) were purified by NiNTA chromatography under denaturing conditions, diluted 20-fold with buffer as described²⁰, and immunopurified by yeast SUMO (SMT3)-specific antibodies or nonspecific immunoglobulin- γ (IgG) antibodies crosslinked to magnetic protein A beads (Dynal). Bound proteins were eluted with 1% SDS. Coomassie-stained bands were excised and analysed by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (Protana). Four peptides of the 45K band matched yeast PCNA (34% coverage). An antibody against yeast PCNA was raised against bacterially expressed His-tagged PCNA from pET28c (Novagen) and affinity purified against PCNA–glutathione S-transferase (GST) expressed from pGEX4 (Amersham Pharmacia). Antibodies against human PCNA (PC10, Abcam), PGK1 (Molecular Probes), ubiquitin, Myc-epitope, UBC9 and CLB2 (all Santa Cruz) were used. Yeast protein extracts were prepared as described⁴⁹. Analytical denaturing NiNTA chromatography/pull-down was done from 0.2 litre of logarithmic cells (*A*_{600nm}) as described²⁰, but at a small scale. For immunoprecipitation, yeast cells were grown to late logarithmic phase and lysed with glass beads in 150 mM NaCl, 50 mM Tris-HCl, pH 7.4, and protease inhibitors. Cell extracts were incubated with detergent (0.2% Triton X-100, 0.1% SDS) and precleared by centrifugation. Proteins were precipitated with Myc-antibody and protein A Sepharose (Amersham Pharmacia). After extensive washing, proteins were eluted with sample buffer containing 8 M urea.

Sensitivity and overexpression assays

For qualitative analysis of sensitivity to MMS or ultraviolet light, stationary-phase cells were either spotted on YPD plates containing MMS or irradiated with ultraviolet light. For quantification, fixed amounts of logarithmic cells were either plated onto YPD plates after incubation with 0.1% MMS for the indicated time, or irradiated with different dosages of ultraviolet light after plating. Colony-forming units were counted after 2–3 days of growth in the dark. Values are averages from two to four experiments. Overexpression analysis was done by spotting serial dilutions onto SC-URA plates containing indicated MMS concentrations.

Cell cycle analysis

For cell-cycle synchronization, logarithmic cells grown at 23 °C were arrested in G1 by 10 μ M α -factor for 2 h, washed, and resuspended in fresh medium. Aliquots were taken at 20-min intervals and analysed by western blotting. Additionally, 200 cells for each time

point were analysed by microscopic inspection for bud formation.

Mammalian cell techniques

Vector for His-tagged PCNA was *pcDNA3.1/GeneStorm* from Invitrogen, and *pMT107* expressing His-tagged UBI was described previously⁵¹. Cells were transfected with Lipofectamine plus reagent (Invitrogen). NiNTA pull-downs were performed as described⁵² but cells were boiled in lysis buffer (100 mM Tris-HCl, pH 7.5, 1% SDS) and lysates were diluted tenfold in PBS before pull-down.

Received 20 May; accepted 10 July 2002; doi:10.1038/nature00991.

- Prakash, S., Sung, P. & Prakash, L. DNA repair genes and proteins of *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **27**, 33–70 (1993).
- Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* (American Society for Microbiology, Washington, 1995).
- Hoeijmakers, J. H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374 (2001).
- Ulrich, H. D. Degradation or maintenance: actions of the ubiquitin system on eukaryotic chromatin. *Eukaryotic Cell* **1**, 1–10 (2002).
- Jentsch, S., McGrath, J. P. & Varshavsky, A. The yeast DNA repair gene *RAD6* encodes a ubiquitin-conjugating enzyme. *Nature* **329**, 131–134 (1987).
- Broomfield, S., Chow, B. L. & Xiao, W. *MMS2*, encoding a ubiquitin-conjugating-enzyme-like protein, is a member of the yeast error-free postreplication repair pathway. *Proc. Natl Acad. Sci. USA* **95**, 5678–5683 (1998).
- Hofmann, R. M. & Pickart, C. M. Noncanonical *MMS2*-encoded ubiquitin-conjugating enzyme functions in assembly of novel polyubiquitin chains for DNA repair. *Cell* **96**, 645–653 (1999).
- Xiao, W. *et al.* Genetic interactions between error-prone and error-free postreplication repair pathways in *Saccharomyces cerevisiae*. *Mutat. Res.* **435**, 1–11 (1999).
- Ulrich, H. D. & Jentsch, S. Two RING finger proteins mediate cooperation between ubiquitin-conjugating enzymes in DNA repair. *EMBO J.* **19**, 3388–3397 (2000).
- Moraes, T. F. *et al.* Crystal structure of the human ubiquitin conjugating enzyme complex, hMms2-hUbc13. *Nature Struct. Biol.* **8**, 669–673 (2001).
- VanDemark, A. P., Hofmann, R. M., Tsui, C., Pickart, C. M. & Wolberger, C. Molecular insights into polyubiquitin chain assembly: crystal structure of the Mms2/Ubc13 heterodimer. *Cell* **105**, 711–720 (2001).
- Bailly, V., Lamb, J., Sung, P., Prakash, S. & Prakash, L. Specific complex formation between yeast *RAD6* and *RAD18* proteins: a potential mechanism for targeting *RAD6* ubiquitin-conjugating activity to DNA damage sites. *Genes Dev.* **8**, 811–820 (1994).
- Pickart, C. M. Ubiquitin in chains. *Trends Biochem. Sci.* **25**, 544–548 (2000).
- Melchior, F. SUMO-nonclassical ubiquitin. *Annu. Rev. Cell Dev. Biol.* **16**, 591–626 (2000).
- Müller, S., Hoege, C., Pyrowolakis, G. & Jentsch, S. SUMO, ubiquitin's mysterious cousin. *Nature Rev. Mol. Cell Biol.* **2**, 202–210 (2001).
- Hochstrasser, M. New structural clues to substrate specificity in the “ubiquitin system”. *Mol. Cell* **9**, 453–454 (2002).
- Tsurimoto, T. PCNA binding proteins. *Frontiers Biosci.* **4**, 849–858 (1999).
- Marbrick, E. The puzzle of PCNA's many partners. *BioEssays* **22**, 997–1006 (2000).
- Bauer, G. A. & Burgers, P. M. Molecular cloning, structure and expression of the yeast proliferating cell nuclear antigen gene. *Nucleic Acids Res.* **18**, 261–265 (1990).
- Johnson, E. S. & Blobel, G. Cell cycle-regulated attachment of the ubiquitin-related protein SUMO to the yeast septins. *J. Cell Biol.* **147**, 981–994 (1999).
- Torres-Ramos, C. A., Yoder, B. L., Burgers, P. M., Prakash, S. & Prakash, L. Requirement of proliferating cell nuclear antigen in *RAD6*-dependent postreplicational DNA repair. *Proc. Natl Acad. Sci. USA* **93**, 9676–9681 (1996).
- Spence, J., Sadis, S., Haas, A. L. & Finley, D. A ubiquitin mutant with specific defects in DNA repair and multiubiquitination. *Mol. Cell Biol.* **15**, 1265–1273 (1995).
- Ayyagari, R., Impellizzeri, K. J., Yoder, B. L., Gary, S. L. & Burgers, P. M. A mutational analysis of the yeast proliferating cell nuclear antigen indicates distinct roles in DNA replication and DNA repair. *Mol. Cell Biol.* **15**, 4420–4429 (1995).
- Joazeiro, C. A. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* **102**, 549–552 (2000).
- Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108**, 345–356 (2002).
- Prakash, L. The *RAD6* gene and protein of *Saccharomyces cerevisiae*. *Ann. NY Acad. Sci.* **726**, 267–273 (1994).
- Johnson, R. E. *et al.* *Saccharomyces cerevisiae* *RAD5*-encoded DNA repair protein contains DNA helicase and zinc-binding sequence motifs and affects the stability of simple repetitive sequences in the genome. *Mol. Cell Biol.* **12**, 3807–3818 (1992).
- Lawrence, C. W. & Christensen, R. B. Metabolic suppressors of trimethoprim and ultraviolet light sensitivities of *Saccharomyces cerevisiae* *rad6* mutants. *J. Bacteriol.* **139**, 866–887 (1979).
- Schiestl, R. H., Prakash, S. & Prakash, L. The *SRS2* suppressor of *rad6* mutations of *Saccharomyces*

- cerevisiae* acts by channeling DNA lesions into the *RAD52* DNA repair pathway. *Genetics* **124**, 817–831 (1990).
- Ulrich, H. D. The *srs2* suppressor of UV sensitivity acts specifically on the *RAD5*- and *MMS2*-dependent branch of the *RAD6* pathway. *Nucleic Acids Res.* **29**, 3487–3494 (2001).
- Broomfield, S. & Xiao, W. Suppression of genetic defects within the *RAD6* pathway by *srs2* is specific for error-free post-replication repair but not for damage-induced mutagenesis. *Nucleic Acids Res.* **30**, 732–739 (2002).
- Zhang, Z., Shibahara, K. & Stillman, B. PCNA connects DNA replication to epigenetic inheritance in yeast. *Nature* **408**, 221–225 (2000).
- Krishna, T. S. *et al.* Crystallization of proliferating cell nuclear antigen (PCNA) from *Saccharomyces cerevisiae*. *J. Mol. Biol.* **241**, 265–268 (1994).
- Amin, N. S. & Holm, C. *In vivo* analysis reveals that the interdomain region of the yeast proliferating cell nuclear antigen is important for DNA replication and DNA repair. *Genetics* **144**, 479–493 (1996).
- Eissenberg, J. C., Ayyagari, R., Gomes, X. V. & Burgers, P. M. Mutations in yeast proliferating cell nuclear antigen define distinct sites for interaction with DNA polymerase δ and DNA polymerase ϵ . *Mol. Cell Biol.* **17**, 6367–6378 (1997).
- Johnson, E. S. & Gupta, A. A. An E3-like factor that promotes SUMO conjugation to the yeast septins. *Cell* **106**, 735–744 (2001).
- Cejka, P., Vondrej, V. & Storchova, Z. Dissection of the functions of the *Saccharomyces cerevisiae* *RAD6* postreplicative repair group in mutagenesis and UV sensitivity. *Genetics* **159**, 953–963 (2001).
- Xiao, W., Chow, B. L., Broomfield, S. & Hanna, M. The *Saccharomyces cerevisiae* *RAD6* group is composed of an error-prone and two error-free postreplication repair pathways. *Genetics* **155**, 1633–1641 (2000).
- Torres-Ramos, C. A., Prakash, S. & Prakash, L. Requirement of *RAD5* and *MMS2* for postreplication repair of UV-damaged DNA in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **22**, 2419–2426 (2002).
- Deng, L. *et al.* Activation of the I κ B kinase complex by TRAF6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain. *Cell* **103**, 351–361 (2000).
- Wang, C. *et al.* TAK1 is a ubiquitin-dependent kinase of MKK and IKK. *Nature* **412**, 346–351 (2001).
- Spence, J. *et al.* Cell cycle-regulated modification of the ribosome by a variant multiubiquitin chain. *Cell* **102**, 67–76 (2000).
- Galan, J. M. & Hagenauer-Tsapis, R. Ubiquitin lys63 is involved in ubiquitination of a yeast plasma membrane protein. *EMBO J.* **16**, 5847–5854 (1997).
- Brusky, J., Zhu, Y. & Xiao, W. *UBC13*, a DNA-damage-inducible gene, is a member of the error-free postreplication repair pathway in *Saccharomyces cerevisiae*. *Curr. Genet.* **37**, 168–174 (2000).
- Tateishi, S., Sakuraba, Y., Masuyama, S., Inoue, H. & Yamaizumi, M. Dysfunction of human *Rad18* results in defective postreplication repair and hypersensitivity to multiple mutagens. *Proc. Natl Acad. Sci. USA* **97**, 7927–7932 (2000).
- Li, Z., Xiao, W., McCormick, J. J. & Maher, V. M. Identification of a protein essential for a major pathway used by human cells to avoid UV-induced DNA damage. *Proc. Natl Acad. Sci. USA* **99**, 4459–4464 (2002).
- Desterro, J. M., Rodriguez, M. S. & Hay, R. T. SUMO-1 modification of I κ B α inhibits NF- κ B activation. *Mol. Cell* **2**, 233–239 (1998).
- Seufert, W., Fletcher, B. & Jentsch, S. Role of a ubiquitin-conjugating enzyme in degradation of S- and M-phase cyclins. *Nature* **373**, 78–81 (1995).
- Knop, M. *et al.* Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines. *Yeast* **15**, 963–972 (1999).
- Li, S. J. & Hochstrasser, M. A new protease required for cell-cycle progression in yeast. *Nature* **398**, 246–251 (1999).
- Treier, M., Staszewski, L. M. & Bohmann, D. Ubiquitin-dependent c-Jun degradation *in vivo* is mediated by the δ domain. *Cell* **78**, 787–798 (1994).
- Müller, S. *et al.* c-Jun and p53 activity is modulated by SUMO-1 modification. *J. Biol. Chem.* **275**, 13321–13329 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank U. Cramer for technical assistance and S. Müller and M. Knop for experimental advice and discussions. We also thank P. Burgers, D. Finley, M. Hochstrasser, M. Knop, S. Müller and B. Stillman for plasmids and strains. S.J. is supported by the Max Planck Society, Deutsche Forschungsgemeinschaft and Fonds der chemischen Industrie.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.J. (e-mail: jentsch@biochem.mpg.de).

Magnetar-like X-ray bursts from an anomalous X-ray pulsar

F. P. Gavril^{*}, V. M. Kaspi^{*†} & P. M. Woods[‡]

^{*} Physics Department, McGill University, Montreal, Quebec H3W 2C4, Canada
[†] Department of Physics and Center for Space Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[‡] Space Science Research Center, National Space Science and Technology Center, Huntsville, Alabama 35805, USA

Anomalous X-ray pulsars (AXPs) are a class of rare X-ray emitting pulsars whose energy source has been perplexing for some 20 years^{1–3}. Unlike other X-ray emitting pulsars, AXPs cannot be powered by rotational energy or by accretion of matter from a binary companion star, hence the designation ‘anomalous’. Many of the rotational and radiative properties of the AXPs are strikingly similar to those of another class of exotic objects, the soft- γ -ray repeaters (SGRs). But the defining property of the SGRs—their low-energy- γ -ray and X-ray bursts—has not hitherto been observed for AXPs. Soft- γ -ray repeaters are thought to be ‘magnetars’, which are young neutron stars whose emission is powered by the decay of an ultra-high magnetic field^{4,5}; the suggestion that AXPs might also be magnetars has been controversial⁶. Here we report two X-ray bursts, with properties similar to those of SGRs, from the direction of the anomalous X-ray pulsar 1E1048.1 – 5937. These events imply a close relationship (perhaps evolutionary) between AXPs and SGRs, with both being magnetars.

SGRs are believed to be magnetars because the high magnetic field provides the torque for their rapid spin-down, as well as the energy to power their bursts and quiescent X-ray emission⁴. AXPs have been suggested to be magnetars, albeit less active, because of their similar spin periods, rates of spin-down and location in the Galactic plane, and their similar (though somewhat softer) X-ray spectra to those of SGRs in quiescence⁷. The physical difference between the two classes is unknown, but, in the magnetar model, is probably related to the magnitude or distribution of the stellar magnetic field. However, the apparent absence of any bursting behaviour in AXPs has led to suggestions that they could be powered, not by magnetism, but by accretion from a disk of material remaining after the birth supernova event⁶. If so, the observational similarities between AXPs and SGRs must be purely coincidental.

A programme to regularly monitor the AXPs using the Proportional Counter Array (PCA)⁸ on board NASA’s Rossi X-ray Timing Explorer (RXTE) satellite was begun in 1996 in order to determine their long-term timing, pulsed flux, and pulse profile stabilities^{9–12}. As part of this programme, motivated by the existence of SGR bursts, we also searched the AXP data for bursts (see Fig. 1 legend for details).

We discovered two highly significant bursts from the direction of AXP 1E1048.1 – 5937. The first (which we call burst 1) occurred during a 3-ks PCA observation obtained on 29 October 2001 with chance probability $P \approx 6 \times 10^{-18}$ after accounting for the number of trials. A second burst (burst 2) was found in a 3-ks observation obtained on 14 November 2001, with analogous probability $P \approx 2 \times 10^{-9}$. No other significant bursts were found towards 1E1048.1 – 5937. The total PCA time searched for bursts towards this source was 380 ks in observations obtained from 1996 to 2002.

The burst profiles are shown in Fig. 1. Both are characterized by fast rises and slow decays (Table 1). Burst 1 appears to have a long, low-level tail that is just above the PCA background as determined by intervals selected before and after the bursts (Fig. 1), whereas burst 2 is much shorter. Both bursts arrived at the peak of the AXP pulse within uncertainties in burst arrival time and definition of

pulse peak. The probability of this occurring by chance is $\sim 1\%$. We note a marginal ($\sim 3\sigma$) increase in the pulsed flux from 1E1048.1 – 5937 that commenced with the observation in which burst 1 was detected, and which lasted ~ 4 weeks.

To determine the bursts’ spectral properties, we first established that neither burst exhibited significant spectral evolution with time by computing hardness ratios (the ratio of 10–60 keV counts to 2–10 keV counts) for the first 0.5-s and subsequent 1.5-s burst intervals. No significant change in hardness was detected, though marginal spectral softening with time was detected after the first 2.5 s of burst 1. Hardness ratios for bursts 1 and 2 for the 1 s after burst onset were 2.8 ± 0.8 and 1.0 ± 0.3 , respectively.

We then fitted the spectra from the first 1 s of each burst to two one-component models—a power law and a black body (Table 1). Continuum models provided an adequate characterization of the burst 2 spectrum but not of the burst 1 spectrum. The spectrum for the 1 s after the burst 1 onset exhibits a feature near 14 keV (Fig. 2). This feature is clear in all binning schemes, and is prominent throughout the first ~ 1 s of the burst. No known PCA instrumental effect produces a feature at this energy (K. Jahoda, personal communication).

Owing to the wide ($\sim 1^\circ$) field of view (FOV) and lack of imaging capabilities of the PCA, we cannot verify that the bursts originated

Table 1 AXP burst timing and spectral properties

	Burst 1	Burst 2
Temporal properties		
Burst day (MJD)	52211	52227
Burst start time (fraction of day, UT)	0.2301949(24)	0.836323379(68)
Burst rise time, t_r (ms)	21^{+9}_{-5}	$5.9^{+2.0}_{-1.2}$
Burst duration, T_{90} (s)	51^{+28}_{-19}	$2.0^{+4.9}_{-0.7}$
Burst phase	-0.018 ± 0.034	0.051 ± 0.032
Fluxes and fluences		
T_{90} fluence (counts)	485 ± 118	101 ± 15
T_{90} fluence ($\times 10^{-10}$ erg cm $^{-2}$)	20.3 ± 4.8	5.3 ± 1.2
1-s fluence (counts)	117 ± 13	69 ± 10
1-s fluence ($\times 10^{-10}$ erg cm $^{-2}$)	$5.9^{+8.6}_{-1.9}$	$4.0^{+3.5}_{-0.8}$
Peak flux for 64 ms ($\times 10^{-10}$ erg s $^{-1}$ cm $^{-2}$)	31^{+45}_{-10}	26^{+23}_{-5}
Peak flux for t_r ($\times 10^{-10}$ erg s $^{-1}$ cm $^{-2}$)	54^{+79}_{-17}	114^{+100}_{-23}
Spectral properties		
Power law		
Power-law index	$0.89^{+1.8}_{-0.71}$	$1.38^{+0.75}_{-0.62}$
Power-law flux ($\times 10^{-10}$ erg s $^{-1}$ cm $^{-2}$)	$2.0^{+8.4}_{-1.8}$	$4.0^{+3.5}_{-0.8}$
Line energy (keV)	13.9 ± 0.9	ND
Line width, σ (keV)	$2.2^{+1.3}_{-1.0}$	ND
Line flux ($\times 10^{-10}$ erg s $^{-1}$ cm $^{-2}$)	$3.9^{+2.2}_{-1.6}$	ND
Reduced χ^2 /degrees of freedom	1.24/15	0.77/5
Black body		
kT (keV)	$3.9^{+3.7}_{-2.7}$	$3.6^{+2.2}_{-1.3}$
Black-body flux ($\times 10^{-10}$ erg s $^{-1}$ cm $^{-2}$)	$2.4^{+5.0}_{-2.1}$	$3.8^{+3.3}_{-1.5}$
Line energy (keV)	$14.2^{+1.1}_{-1.2}$	ND
Line width (keV)	$2.1^{+1.5}_{-1.3}$	ND
Line flux ($\times 10^{-10}$ erg s $^{-1}$ cm $^{-2}$)	$3.7^{+2.2}_{-1.9}$	ND
Reduced χ^2 /degrees of freedom	1.23/15	1.66/5

The uncertainty on the burst start time is the burst rise time t_r , and is given in parentheses as the uncertainty in the last digits shown. The burst rise times were determined by a maximum-likelihood fit to the unbinned data using a piecewise function having a linear rise and exponential decay. The burst duration, T_{90} , is the interval between when 5% and 95% of the total 2–20 keV burst fluence was received. The background regions used for both the duration and spectral analyses are shown in Fig. 1. Burst phase is defined such that the peak of the periodic pulsation is at phase 0/1. All fluences and fluxes are in the 2–20 keV range. T_{90} fluences in c.g.s. units were calculated assuming a power-law spectral model and spectral grouping that demanded a minimum of 20 counts per spectral bin. The 1-s fluences in c.g.s. units correspond to the fluxes found in the spectral modelling. The spectral rebinning method used in all spectral modelling for burst 1 was to group the 256 PCA channels by a factor of 4, whereas for burst 2, we demanded at least 20 counts per spectral bin. Peak fluxes on the short timescales were determined by scaling the 1-s fluxes by number of counts. For all spectral fits, the equivalent neutral hydrogen column density was held fixed at 1.2×10^{22} cm $^{-2}$, the value determined from recent XMM observations²⁵. Spectral fits were determined for the first 1 s after the burst start times. Spectral modelling was done using photons in the 2–40 keV range. Response matrices were created using the FTOOL pcarsp (http://heasarc.gsfc.nasa.gov/docs/xte/recipes/pca_response.html). Uncertainties in the table are 68% confidence intervals, except for those reported for the c.g.s.-unit fluences and fluxes, as well as the spectral model parameters, for which we report 90% confidence intervals. ND, not determined.

from the location of the AXP. The low peak X-ray fluxes of the events (Table 1) preclude determining the source's location using data from other, better imaging instruments that were contemporaneously observing the X-ray sky, such as the RXTE All Sky Monitor, or the Wide Field Camera on board the BeppoSAX satellite. We must therefore consider other possible origins for the bursts before concluding they were from the AXP.

The bursts' short rise times (Table 1) require emission regions of less than a few thousand kilometres, implying a compact object origin. So-called type I X-ray bursts are well-studied phenomena that result from unstable helium burning just below the surface of a weakly magnetized neutron star that is accreting material in a low-mass X-ray binary (LMXB)¹³. However, type I bursts from an LMXB in the same FOV as 1E1048.1 – 5937 are unlikely to explain our observed bursts because: (1) the burst rise times are much shorter than those of type I bursts; (2) the burst spectra are much harder than those of type I bursts; (3) burst 2 shows no evidence for spectral softening with time, and no type I burst has ever exhibited a spectral feature like the one detected in burst 1; (4) the bursts are extremely faint, implying a source location well outside the Milky Way for type I burst luminosities; and (5) there are no known LMXBs in the FOV¹⁴. Type II X-ray bursts¹³ are a much rarer and less well understood phenomenon observed thus far in only two sources, both accreting binaries. The bursts that we have observed are unlikely to be type II bursts from an unknown X-ray binary in the PCA FOV for the following reasons: first, because of the rarity of such events; second, type II bursts have longer rise times than do our bursts; and third, no type II burst has exhibited a spectral feature like that seen in burst 1.

Classical γ -ray bursts (GRBs) sometimes show prompt X-ray emission that can have temporal and spectral signatures similar to those we have observed¹⁵. However, the likelihood of two GRBs occurring within 1° of each other is small, and GRBs are not known

to repeat. We make the conservative assumption of GRB spectral model parameters that result in low γ -ray fluxes, and extrapolate the GRB rate¹⁶ as measured with the Burst and Transient Source Experiment (BATSE)¹⁷ assuming homogeneity below the BATSE threshold. We then estimate a probability of $\sim 9 \times 10^{-5}$ that these events are unrelated GRBs that occurred by chance in the same RXTE FOV during our 1E1048.1 – 5937 monitoring observations (conservatively neglecting that they occurred within two weeks of each other).

The observed burst properties are in many ways similar to those seen from SGRs¹⁸. The fast rise and slow decay profiles are consistent with SGR time histories, as are the burst durations (neglecting the long, low-level tail of burst 1). Both AXP and SGR bursts are spectrally much harder than is their quiescent pulsed emission. The burst peak fluxes and fluences fall within the range seen for SGRs, and the spectrum of burst 2 is consistent with SGR burst spectra of comparable fluence. Burst 1 has characteristics unlike those of nearly all SGR bursts, specifically its long tail and spectral feature (Fig. 2). However, we note that a single event from SGR1900 + 14 was shown^{19,20} to possess each of these properties. The marginal increase in the pulsed fraction that we observed at the burst epochs is consistent with SGR pulsed flux increases seen during bursting episodes²¹. Finally, the fact that in spite of several years of monitoring, the only two bursts detected occurred within two weeks of each other suggests episodic bursting activity, the hallmark of SGRs. Thus, the characteristics of these events match the burst properties of SGRs far better than any other known burst phenomenon.

In the magnetar model for SGRs^{4,7}, bursts are a result of sudden crustal yields owing to stress from the outward diffusion of the huge internal magnetic field. Such yields cause crust shears which twist the external magnetic field, releasing energy. Thompson and Duncan⁷, who suggested that AXPs are also magnetars, also predicted

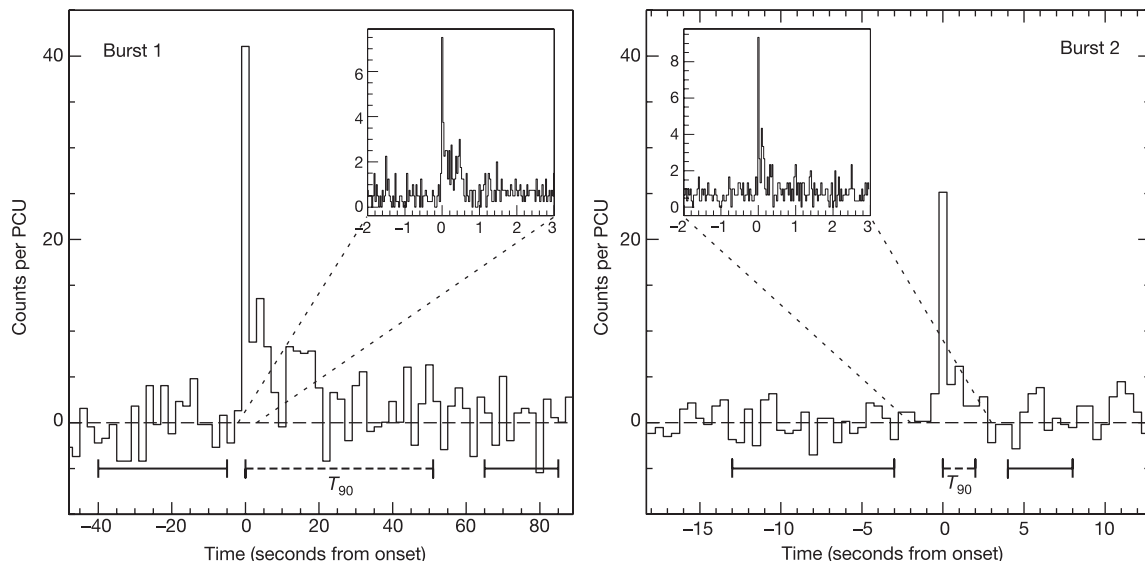


Figure 1 Light curves for the observed bursts. The RXTE AXP data set consists of short (~ 3 ks) snapshots, as well as longer archival observations, all taken in the PCA 'GoodXenonWithPropane' mode, which records photon arrival times with $1\text{-}\mu\text{s}$ resolution, and bins photon energies into 256 channels. Time series were initially created with 31.25-ms resolution from photons having energies in the range 2–20 keV for each PCA Proportional Counter Unit (PCU) separately, using all xenon layers. Photon arrival times at each epoch were adjusted to the Solar System barycentre. The resulting time series were searched for significant excursions from the mean count rate by comparing each time bin value with a windowed 7-s running mean. Bursts were identified assuming poissonian statistics, and by combining probabilities from the separate PCUs. Left, background-

subtracted 2–20 keV light curves for burst 1, displayed with 2-s time resolution. The solid horizontal lines before and after the bursts are the boundaries of the pre- and post-background intervals used for calculating T_{90} and for spectral modelling. The T_{90} interval is shown as a horizontal dashed line. Right, same but for burst 2, and with 0.5-s time resolution in the main panel. The insets show the peak of each burst with 31.25-ms time resolution. We verified that there was no significantly enhanced signal from PCA events not flagged as 'good' at the time of the bursts (such as those that do not enter through the PCA aperture) in the RXTE "Standard 1" event files. We also verified that both events were clearly detected in all operational PCUs. Hence the events are unlikely to be instrumental in origin.

that X-ray bursts should eventually be seen from them. By contrast, SGR-like bursts are not expected in any AXP accretion model, whether binary or isolated fall-back disk.

The large 14-keV spectral line in burst 1 is intriguing. An electron cyclotron feature at this energy E implies a magnetic field of $B = 2\pi mcE/h \approx 1.2 \times 10^{12}$ G (where m is the electron mass, c is the speed of light, h is Planck's constant, and e is the electron charge), whereas a proton cyclotron feature implies $B \approx 2.4 \times 10^{15}$ G. The former field strength is significantly lower than is implied from the source's spin-down, and is typical of conventional young neutron stars, rather than magnetars; the latter field strength is higher than is implied by the spin-down, yet reasonable for the magnetar model as the spin-down torque is sensitive only to the dipolar component of the magnetic field.

Why do the burst rates of AXPs and SGRs differ so markedly, in spite of their common magnetar nature? One possibility is that AXP internal magnetic fields are much larger than those of SGRs; if so, AXP crusts could undergo plastic deformation rather than brittle fracturing⁷. However, this is opposite to what is inferred from the

spin-down rates of the two classes, suggesting that spin down is an unreliable internal field indicator. This could help reconcile the contrasting radiative properties of AXPs and apparently high-magnetic-field radio pulsars²². It also suggests that AXPs are SGR progenitors, with bursting behaviour starting as the field decays. This is consistent with the smaller AXP ages implied by their more numerous associations with supernova remnants²³, but does not explain why AXPs and SGRs have similar spin period distributions, as AXPs spin down as they age²⁴. This aspect of magnetar physics remains a puzzle. □

Received 11 June; accepted 19 July 2002; doi:10.1038/nature01011.

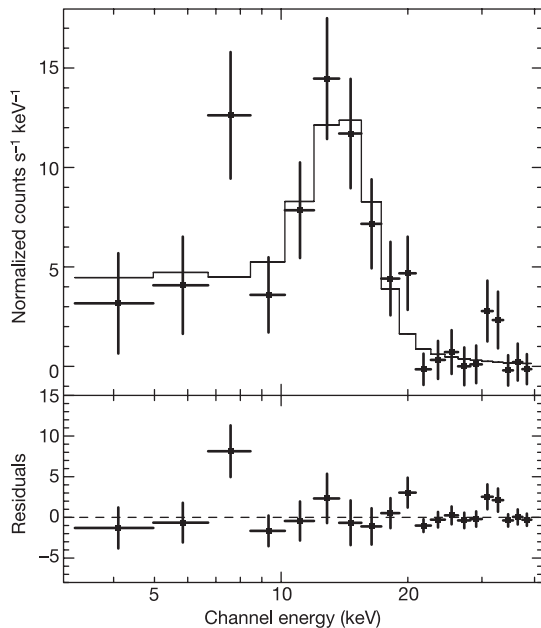


Figure 2 X-ray spectrum in the 2–40 keV range for the 1 s after the onset of burst 1. For all spectral analyses, we first created spectral files having 256 channels across the full PCA energy range (~ 0.2 –60 keV), although subsequent factor of 4 rebinning was necessary because of the paucity of burst counts. The burst and background intervals were used as input to the X-ray spectral fitting package XSPEC²⁶ v1.1.0 (<http://xspec.gsfc.nasa.gov>). The spectrum of the first 1 s after burst 1 onset is not well characterized by any continuum model. The best fit power-law plus gaussian line model is shown as a solid line. The F -test shows that the addition of a line of arbitrary energy, width and normalization to a simple power-law model improves the fit significantly, with a chance probability of this occurring of 0.0032. Monte Carlo simulations in XSPEC were done to verify this conclusion: 10,000 simulations of similar data sets were produced assuming a simple power-law energy distribution, then fitted with a power law plus gaussian line of arbitrary energy, width and normalization. This procedure is conservative, as it ignores the fact that the observed large line has flux comparable to the measured continuum. In 10,000 trials, we found 1 trial with the same or smaller chance occurrence probability as judged by the F -test, indicating that the probability of the line we observed being due to chance is <0.0001 . We repeated this procedure for data having a black-body spectrum, with similar results, namely, the probability of the line being due to chance is <0.0008 . The spectrum also shows possible additional features at ~ 7 keV and ~ 30 keV (suggestive of lines at multiples of 1, 2 and 4 times ~ 7 keV). These additional features are not apparent in all binning schemes, and are not statistically significant.

1. Fahlman, G. G. & Gregory, P. C. An X-ray pulsar in SNR G109.1–1.0. *Nature* **293**, 202–204 (1981).
2. van Paradijs, J., Taam, R. E. & van den Heuvel, E. P. J. On the nature of the ‘anomalous’ 6-s X-ray pulsars. *Astron. Astrophys.* **299**, L41–L44 (1995).
3. Mereghetti, S. & Stella, L. The very low mass X-ray binary pulsars: A new class of sources? *Astrophys. J.* **442**, L17–L20 (1995).
4. Thompson, C. & Duncan, R. C. The soft gamma repeaters as very strongly magnetized neutron stars—I. Radiative mechanism for outbursts. *Mon. Not. R. Astron. Soc.* **275**, 255–300 (1995).
5. Kouveliotou, C. et al. An X-ray pulsar with a superstrong magnetic field in the soft γ -ray repeater SGR 1806–20. *Nature* **393**, 235–237 (1998).
6. Chatterjee, P., Hernquist, L. & Narayan, R. An accretion model for anomalous X-ray pulsars. *Astrophys. J.* **534**, 373–379 (2000).
7. Thompson, C. & Duncan, R. C. The soft gamma repeaters as very strongly magnetized neutron stars—II. Quiescent neutrino, x-ray, and Alfvén wave emission. *Astrophys. J.* **473**, 322–342 (1996).
8. Jahoda, K. et al. In-orbit performance and calibration of the Rossi X-ray Timing Explorer (RXTE) proportional counter array (PCA). *Proc. SPIE* **2808**, 59–70 (1996).
9. Kaspi, V. M., Chakrabarty, D. & Steinberger, J. Phase-coherent timing of two anomalous X-ray pulsars. *Astrophys. J.* **525**, L33–L36 (1999).
10. Kaspi, V. M., Lackey, J. R. & Chakrabarty, D. A glitch in an anomalous X-ray pulsar. *Astrophys. J.* **537**, L31–L34 (2000).
11. Kaspi, V. M., Gavril, P., Chakrabarty, D., Lackey, J. R. & Munro, M. P. Long-term RXTE monitoring of the anomalous X-ray pulsar 1E 1048.1–5973. *Astrophys. J.* **558**, 253–262 (2001).
12. Gavril, P. & Kaspi, V. M. Long-Term Rossi X-ray Timing Explorer monitoring of anomalous X-ray pulsars. *Astrophys. J.* **567**, 1067–1076 (2002).
13. Lewin, W. H. G., van Paradijs, J. & Taam, R. E. *X-ray Binaries* (eds Lewin, W. H. G. & van Paradijs, J. & van den Heuvel, E. P. J.) 175–232 (Cambridge Univ. Press, Cambridge, 1995).
14. Liu, Q. Z., van Paradijs, J. & van den Heuvel, E. P. J. A catalogue of low-mass X-ray binaries. *Astron. Astrophys.* **368**, 1021–1054 (2001).
15. Heise, J., in’t Zand, J., Kippen, M. & Woods, P. X-ray flashes and X-ray rich gamma-ray bursts. *Mem. Soc. Astron. Ital.* (in the press); preprint astro-ph/0111246 at (<http://xxx.lanl.gov>) (2001).
16. Stern, B. E., Tikhomirova, Y., Kompaneets, D., Svensson, R. & Poutanen, J. An off-line scan of the BATSE daily records and a large uniform sample of gamma-ray bursts. *Astrophys. J.* **563**, 80–94 (2001).
17. Fishman, G. J. et al. Overview of observations from BATSE on the Compton observatory. *Astron. Astrophys. Suppl.* **97**, 17–20 (1993).
18. Göğüş, E. et al. Temporal and spectral characteristics of short bursts from the soft gamma repeaters 1806–20 and 1900 + 14. *Astrophys. J.* **558**, 228–236 (2001).
19. Ibrahim, A. I. et al. An unusual burst from soft gamma repeater SGR 1900 + 14: Comparisons with giant flares and implications for the magnetar model. *Astrophys. J.* **558**, 237–252 (2001).
20. Strohmayer, T. E. & Ibrahim, A. I. Discovery of a 6.4 keV emission line in a burst from SGR 1900 + 14. *Astrophys. J.* **537**, L111–L114 (2000).
21. Woods, P. M. et al. Evidence for a sudden magnetic field reconfiguration in soft gamma repeater 1900 + 14. *Astrophys. J.* **552**, 748–755 (2001).
22. Pivovarov, M., Kaspi, V. M. & Camilo, F. X-ray observations of the high magnetic field radio pulsar PSR J1814–1744. *Astrophys. J.* **535**, 379–384 (2000).
23. Gotthelf, E. V., Vasisth, G. & Dotani, T. On the spin history of the X-ray pulsar in Kes 73: Further evidence for an ultramagnetized neutron star. *Astrophys. J.* **522**, L49–L52 (1999).
24. Gaensler, B. M., Slane, P. O., Gotthelf, E. V. & Vasisth, G. Anomalous X-ray pulsars and soft gamma-ray repeaters in supernova remnants. *Astrophys. J.* **559**, 963–972 (2001).
25. Tiengo, A., Göhler, E., Staubert, R. & Mereghetti, S. The anomalous X-ray pulsar 1E 1048.1–5973: Phase resolved spectroscopy with the XMM-Newton satellite. *Astron. Astrophys.* **383**, 182–187 (2002).
26. Arnaud, K. A. *Astronomical Data Analysis Software and Systems V* (eds Jacoby, G. & Barnes, J.) 17 (ASP, San Francisco, 1996).

Acknowledgements

P.M.W. is affiliated to the Universities Space Research Association. We thank E. Kuulkers, K. Hurley, K. Jahoda, C. Kouveliotou, W. Lewin, M. Lyutikov, M. Munro, D. Psaltis, S. Ransom, M. Roberts, D. Smith and C. Thompson for discussions. This work was supported in part by a NASA LTSA grant, an NSERC research grant, and an NSF career award. V.M.K. is a Canada Research Chair and an Alfred P. Sloan Fellow. This work made use of data obtained through the High Energy Astrophysics Science Archive Research Center Online Service, provided by the NASA/Goddard Space Flight Center.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.M.K. (e-mail: vkaspi@physics.mcgill.ca).

Simultaneous micromanipulation in multiple planes using a self-reconstructing light beam

V. Garcés-Chávez, D. McGloin, H. Melville, W. Sibbett & K. Dholakia

School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, Fife KY16 9SS, UK

Optical tweezers¹ are commonly used for manipulating microscopic particles, with applications in cell manipulation², colloid research^{3–5}, manipulation of micromachines⁶ and studies of the properties of light beams⁷. Such tweezers work by the transfer of momentum from a tightly focused laser to the particle, which refracts and scatters the light and distorts the profile of the beam. The forces produced by this process cause the particle to be trapped near the beam focus. Conventional tweezers use gaussian light beams, which cannot trap particles in multiple locations more than a few micrometres apart in the axial direction, because of beam distortion by the particle and subsequent strong divergence from the focal plane. Bessel beams^{8,9}, however, do not diverge and, furthermore, if part of the beam is obstructed or distorted the beam reconstructs itself after a characteristic propagation distance¹⁰. Here we show how this reconstructive property may be utilized within optical tweezers to trap particles in multiple, spatially separated sample cells with a single beam. Owing to the diffractionless nature of the Bessel beam, secondary trapped particles can reside in a second sample cell far removed (~3 mm) from the first cell. Such tweezers could be used for the simultaneous study of identically prepared ensembles of colloids and biological matter, and potentially offer enhanced control of 'lab-on-a-chip' and optically driven microstructures.

Any light beam can be thought of as a superposition of plane waves. As the waves propagate a distance Δz , they undergo a phase shift $k_z \Delta z$, where k_z is the wavevector of the beam in the z direction. In most cases, each plane wave component suffers a different phase shift, and so the resulting beam—the interference pattern of the plane waves—changes shape. There do exist, however, special beams where the phase shift is the same for every plane wave component. These beams do not change shape on propagation, and therefore may be considered diffraction free. For this to happen, the phase shift $k_z \Delta z$ must be the same for all the plane wave components, which can happen if all the waves propagate on a cone. Such a cone can readily be created by illuminating a thin annulus in the front focal plane of a lens. So we can see how such a beam may also be able to reconstruct itself: if part of the beam is blocked, then a shadow is cast into the beam, but the plane waves on the cone that pass the obstruction can reform the beam at a point just beyond the length of the shadow. Such self-reconstruction properties are characteristic of a Bessel beam.

A practical realization of a Bessel beam¹¹ can be produced by illuminating a conical shaped optical element, called an axicon, with

a gaussian beam¹². The beam produced is a close approximation to a Bessel beam over a characteristic propagation distance. The central maximum propagates for several Rayleigh ranges without appreciable divergence, and thus approximates a rod of light. The outer rings of the Bessel beam act to replenish the central maximum and prevent it from spreading. This replenishment also allows the Bessel beam to reconstruct itself if blocked¹³. If a beam has an object, of radius r_{ob} , placed at its centre, then the object will cast a shadow of length l_s into the beam:

$$l_s \approx \frac{r_{\text{ob}} k}{k_r} \quad (1)$$

where k_r is the radial wavevector of the beam, with a wavevector $k = (k_r^2 + k_z^2)^{1/2}$.

After this distance the beam will have reformed, and will continue to propagate without diffraction over its remaining propagation distance. In Fig. 1 we show a numerical simulation of a zeroth-order obstructed Bessel beam. We could also use high-order Bessel beams, which have an intensity minimum on axis—such beams also reconstruct when blocked. In addition, there are other families of diffractionless beams, such as the Mathieu beams¹⁴, which have properties comparable to (but distinct from) Bessel beams and that can be used similarly to the work presented here. We are also able to generate other patterns of diffractionless beams, notably by creating interference patterns (between two other non-diffracting beams of opposite helicity) that have a number of spots equal to twice the azimuthal index of the interfering beams (D.M., V.G.-C. and K.D., manuscript in preparation). By using spatial light modulators¹⁵, it is possible to create arrays of Bessel beams in any desired pattern, adding a greater generality to our technique.

Bessel beams can be used as simple two-dimensional optical tweezers¹⁶. As they are propagation invariant, they act as a rod of light and have no confining force in the beam propagation direction. We are able to create tweezers in two differing geometries. Standard (downward) tweezers push (via radiation pressure) the samples down against the microscope slide. Inverted (upward) tweezers are used to levitate, align, stack and guide particles, again

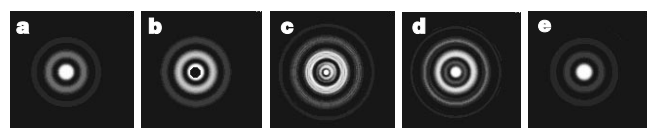


Figure 1 Beam propagation simulation. Numerical simulation of the propagation of a Bessel beam that has its central maximum blocked. **a**, Bessel beam before the obstruction—a set of concentric rings surrounding a bright central spot. **b**, Obstruction placed over the central spot. **c–e**, The beam at various positions beyond the obstruction. The beam is seen to distort, and then reform in **e**. Boxes are 0.75 mm squares.

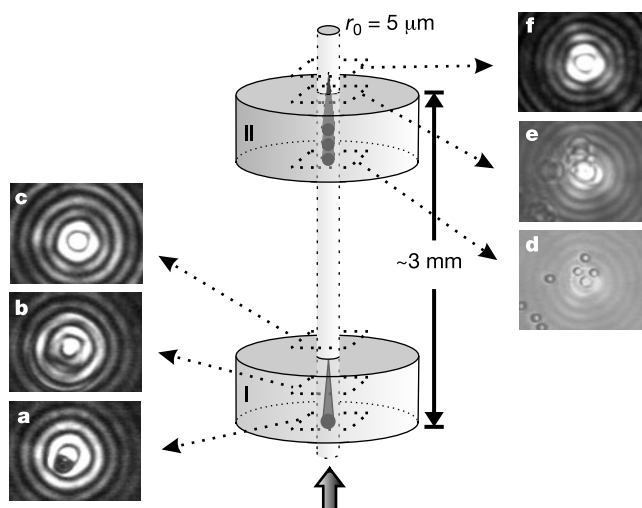


Figure 2 Inverted tweezers experimental set-up. Main figure, set-up for manipulation of particles that have large spatial separations. The cells (I and II) are 3 mm apart, and 100 μm deep. **a–f**, Frames from a video taken of objects captured by the Bessel beam; control spot radius $r_0 = 5 \mu\text{m}$. **a**, A hollow sphere ($n < 1$) of $\sim 5 \mu\text{m}$ diameter is trapped in cell I between the central spot and the first ring of the Bessel beam. Note that n is the relative refractive index between the particle and the suspending medium. **b**, The beam a short distance above **a**. Here the beam has been distorted by the particle. **c**, Some small distance above the first sample cell, the beam has reformed and is no longer distorted. **d**, The beam enters the cell II, and is able to stack three $5\text{-}\mu\text{m}$ -diameter solid silica spheres. **e**, **f**, The beam profile above the stack of particles. The beam has reformed once more.

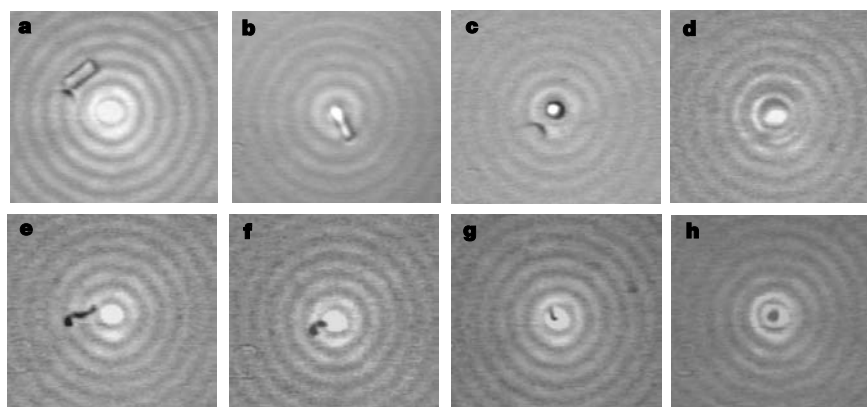


Figure 3 Alignment of glass rods and chromosomes. Sequence of frames taken from a video of optical manipulation in two separated cells (Fig. 2). Frames **a–c** and **e–h** show particles being attracted to the central spot of the Bessel beam, and then aligning themselves with the beam. The lower cell (top row) contains a glass rod of $\sim 8\ \mu\text{m}$ length; the upper cell (bottom row) contains a Muntjac deer chromosome. **a**, The profile of the

beam entering the bottom cell. **b**, the glass rod trapped in the central maximum of the Bessel beam; it is eventually rotationally aligned along the beam direction (**c**). **d**, The distorted beam after passing the confined rod. The beam continues to propagate to the top cell (**e**), where the chromosome is evident. **f–h**, The subsequent trapping and rotational alignment of the chromosome.

making use of radiation pressure in the direction of the beam. For the observations reported here we use both geometries, with a basic experimental set-up described previously¹⁶. An axicon is illuminated with a power of up to 1 W from a 1,064-nm-wavelength laser. Telescoping optics before the axicon allow us to vary the propagation length of the Bessel beam. A variable-magnification telescopic system is placed after the axicon, allowing us to reduce the size of the central maximum. In this manner we can readily adjust the parameters of the Bessel beam that is incident on the sample slide. The maximum power at the sample plane was about 700 mW, which is spread evenly among the rings, with a maximum of about 35 mW found in the central spot of the Bessel beam. The power measured after each reconstruction of the beam is approximately the same. We are able to control the power in the rings by changing the number of rings in the Bessel beam. Decreasing the number of rings will increase the power in each ring (and the central spot), but this comes at the expense of a decrease in propagation distance. A microscope objective and CCD (charge-coupled device) camera are used to observe the trapped particles on a video monitor.

We discuss our experimental results on Bessel beam optical tweezers in both the inverted and the standard geometries. Figure 2 shows results, in the inverted geometry, using two sample cells 3 mm apart. Low-refractive-index (with respect to the suspending medium) particles are in the bottom cell, and 5- μm -diameter silica spheres are in the top cell. Each cell was $\sim 100\ \mu\text{m}$ deep. We used a Bessel beam with a central radius of $\sim 5\ \mu\text{m}$ and a propagation distance of about 4 mm. The beam profile was distorted by the low-refractive-index particle, and reconstructed 100 μm beyond it. Using exactly the same beam and after passing the bottom cell, we have stacked up to three 5- μm spheres above one another, in a one-dimensional array, in the top cell, and manipulated this whole particle chain while still manipulating in unison the low-refractive-index particle trapped in the first cell. We note that the separation distance between the cells (3 mm) is ~ 40 times the Rayleigh range of a gaussian beam with a 5- μm beam waist.

We also observe the ability to use these tweezers to align rod-shaped particles in multiple cells. Many biological specimens have such forms, and by using Bessel beam tweezers the rods can be aligned along the propagation direction of the beam¹⁶. This is achieved using a single stationary beam—when trapped, the particles are held simultaneously. (This is in contrast to the single particle that could be aligned with a gaussian beam.) They can then be moved in unison by translation of the sample cells. Here we use a Bessel beam with a propagation distance of $\sim 10\ \text{mm}$, using the same cell set-up (Fig. 2). In the bottom cell, we have fragments of

glass rod that are approximately 8 μm in length, and in the top cell we have Muntjac deer chromosomes. Figure 3 shows the alignment of the rods and chromosomes in their respective cells.

Using the downward tweezers geometry, we observed tweezer action in three separate cells—again each $\sim 100\ \mu\text{m}$ in depth, with $\sim 10\text{-}\mu\text{m}$ cover glass slips as separators—stacked one on top of the other. The top cell contains 5- μm -diameter solid silica spheres. In the second cell are 2–20- μm -diameter hollow silica spheres, and in the third cell are birefringent (calcite, $\sim 5 \times 4\ \mu\text{m}$) particles. We are able to trap particles in all three traps simultaneously and manipulate them.

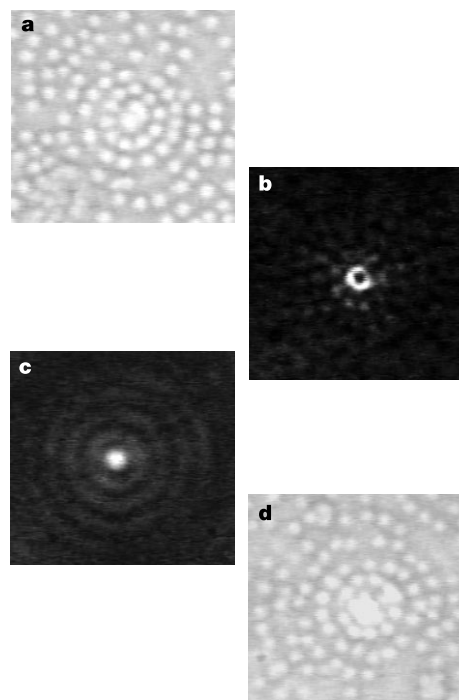


Figure 4 Arrays of 1 μm spheres. **a–d**, Successive frames taken from a video, with the trapping beam passing through two cells containing 1- μm spheres in water. The set-up uses the standard, downward geometry, tweezers. The two cells are separated by 100 μm . **a**, The bottom of the upper cell. The beam profile is distorted (**b**), and subsequently self-reconstructs (**c**) below the first cell. **d**, The pattern created in the lower cell. Frames **a** and **d** appear bright as an extra white-light source is used to see the sample. Notice that the beam is able to form a two-dimensional array in each cell.

Our single laser beam can trap particles of varying refractive index, which cannot be done with a single gaussian beam. Hollow spheres (relative refractive index $n < 1$) cannot be trapped by a normal gaussian beam, because regions of high intensity repel such particles. By contrast, we are able to trap such particles with our Bessel beam—between the bright rings, in the dark regions when the sphere diameter exceeds the bright ring spacing. We also note that we are not only able to trap the birefringent particle at the bottom of the third cell, but we can rotate it as well, using a circularly polarized Bessel beam. We thus impart angular momentum to the particle¹⁷. The different types of particle have specific reconstruction distances, indicating that the reconstruction distance is, at least in part, dependent on the refractive index of the particle. Owing to the differing distortion and shadowing of a Bessel beam by different types of particles, there is the prospect of using the self-reconstructing beam as a means to characterize particles or cells. The number of particles we can trap is limited, but only by the available laser power and the propagation distance of the Bessel beam. The size of the obstruction ultimately represents a limitation, because if the entire beam is blocked then we could not expect any subsequent reconstruction. Particles that we can trap with tweezers are generally at least partially transparent, and hence even if the sample cell is filled with particles the transmitted beam is still able to reconstruct. We show in Fig. 4 a dense sample (~20% of the sample area) of 1- μm -diameter silica spheres that cover the observable Bessel beam in the cell. The particles are trapped in the rings of the Bessel beam, forming a two-dimensional planar array. We then see an image from a sample cell 100 μm below the top cell also filled with 1- μm -diameter silica spheres. We see that these particles can be manipulated, and also that the Bessel beam is able to reform. The power in the reformed Bessel beam is ~90% (630 mW) of the power of the beam incident on the first cell. Thus we have created multiple planes of 2D arrays where the planes of the array are separated by the depth of the sample cell.

As mentioned above, we have the ability to rotate birefringent particles by transference of spin angular momentum from the beam to the particle in a multi-cell system. A more interesting extension of this work is to interfering Bessel beams. Recent work^{18,19} has shown that by interfering high-order Laguerre–Gaussian beams it is possible to create interferometric patterns that consist of arrays of spots. These can be used to rotate particles, which need not be birefringent, in a controlled fashion, or alternatively to create three-dimensional structures. We can also create such patterns with Bessel beams, and have used these to rotate particles (D.M., V.G.-C. and K.D., manuscript in preparation). This opens up the possibility of the controlled rotation of systems (that is, micromachines), or the guidance and alignment of streams of particles, or cold atoms, over extended distances in unison. Using Bessel beams for rotational studies of particles may be of particular practical interest. We are able to rotate particles in multiple cells, rotating many objects simultaneously, thereby opening up the possibility of arrays of micromachines. Such beams could act on more complex microstructures. For example, imagine two micro-cogs, attached by an axle rod. As the Bessel beam reconstructs itself, we could use the same beam to drive both cogs simultaneously and with the same rotation speed.

Self-reconstructing optical tweezers could also be applied to other avenues of research. For example, it may be possible to conduct simultaneous studies of identically prepared samples, in different cells, with each cell having a different ambient condition—such as temperature. We have shown how arrays of particles may be created, which could have applications in colloid physics and in bioengineering. This work could be readily extended to the creation of large three-dimensional arrays. Conventionally this is difficult. Recent work^{15,20} has shown that by using holographic beam shaping techniques, two-dimensional planar arrays can be created. These can also be dynamically controlled. An array of such dynamic

tweezers, making use of reconstructing Bessel beams, would represent a controllable three-dimensional structure, and could have significant application in, for instance, the control of arrays of lab-on-a-chip microstructures and micromachines. □

Received 10 May; accepted 23 July 2002; doi:10.1038/nature01007.

1. Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E. & Chu, S. Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.* **11**, 288–290 (1986).
2. Smith, S. B., Cui, Y. & Bustamante, C. Overstretching B-DNA: The elastic response of individual double stranded and single stranded DNA molecules. *Science* **271**, 795–799 (1996).
3. Crocker, J. C. & Grier, D. G. Methods of digital video microscopy for colloidal studies. *J. Colloid Interf. Sci.* **179**, 298–310 (1996).
4. Crocker, J. C., Matteo, J. A., Dinsmore, A. D. & Yodh, A. G. Entropic attraction and repulsion in binary colloids probed with a line optical tweezer. *Phys. Rev. Lett.* **82**, 4352–4355 (1999).
5. Larsen, A. E. & Grier, D. G. Like charge attractions in metastable colloidal crystallites. *Nature* **385**, 230–233 (1997).
6. Friese, M. E. J., Rubinsztajn-Dunlop, H., Gold, J., Hagberg, P. & Hanstorp, D. Optically driven micromachine elements. *Appl. Phys. Lett.* **78**, 547–549 (2001).
7. Volke-Sepulveda, K., Garcés-Chávez, V., Chávez-Cerda, S., Arlt, J. & Dholakia, K. Orbital angular momentum of a high-order Bessel light beam. *J. Opt. B* **4**, 582–589 (2002).
8. Durnin, J., Miceli, J. J. Jr & Eberly, J. H. Diffraction-free beams. *Phys. Rev. Lett.* **58**, 1499–1501 (1987).
9. McQueen, C. A., Arlt, J. & Dholakia, K. An experiment to study a “nondiffracting” light beam. *Am. J. Phys.* **67**, 912–915 (1999).
10. Bouchal, Z., Wagner, J. & Chlup, M. Self-reconstruction of a distorted nondiffracting beam. *Opt. Commun.* **151**, 207–211 (1998).
11. Durnin, J. Exact solutions for nondiffracting beams. I. The scalar theory. *J. Opt. Soc. Am. A* **4**, 651–654 (1987).
12. Herman, R. M. & Wiggins, T. A. Production and uses of diffractionless beams. *J. Opt. Soc. Am. A* **8**, 932–942 (1991).
13. MacDonald, R. P., Boothroyd, S. A., Okamoto, T., Chrostowski, J. & Syrett, B. A. Interboard optical data distribution by Bessel beam shadowing. *Opt. Commun.* **122**, 169–177 (1996).
14. Chávez-Cerda, S. et al. Holographic generation and orbital angular momentum of high-order Mathieu beams. *J. Opt. B* **4**, S52–S57 (2002).
15. Curtis, J. E., Koss, B. A. & Grier, D. G. Dynamic holographic optical tweezers. *Opt. Commun.* **207**, 169–175 (2002).
16. Arlt, J., Garcés-Chávez, V., Sibbett, W. & Dholakia, K. Optical micro-manipulation using a Bessel light beam. *Opt. Commun.* **197**, 239–245 (2001).
17. Friese, M. E. J., Nieminen, T. A., Heckenberg, N. R. & Rubinsztajn-Dunlop, H. Alignment or spinning of laser-trapped microscopic waveplates. *Nature* **394**, 348–350 (1998).
18. Paterson, L. et al. Controlled rotation of optically trapped microscopic particles. *Science* **292**, 912–914 (2001).
19. MacDonald, M. P. et al. Creation and manipulation of three-dimensional optically trapped structures. *Science* **296**, 1101–1103 (2002).
20. Korda, P., Spalding, G. C., Dufresne, E. R. & Grier, D. G. Nanofabrication with holographic optical tweezers. *Rev. Sci. Instrum.* **73**, 1956–1957 (2002).

Acknowledgements

We thank G. Spalding for discussions. This work was supported by the Leverhulme Trust, the UK Engineering and Physical Sciences Research Council and the Medical Research Council.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.D. (e-mail: kd1@st-and.ac.uk).

Mechanical milling assisted by electrical discharge

A. Calka & D. Wexler

Faculty of Engineering, University of Wollongong, Wollongong, New South Wales 2522, Australia

Mechanical milling is an effective technique for the preparation of fine metallic and ceramic powders and can also be used to drive a wide range of chemical reactions. Milling devices include planetary machines, attritors and vibrational mills; products include amorphous, nanocrystalline and quasicrystalline materials, supersaturated solid solutions, reduced minerals,

high-surface-area catalysts and reactive chemicals^{1–3}. During milling, solid–solid, solid–liquid and solid–gas reactions are initiated through repeated deformation and fracture of powder particles. A separate materials synthesis and processing technique involves reacting a material in a gas atmosphere under an electrical discharge^{4–7}. Here we show that the application of low-current, high-voltage electrical impulses during milling can result in both faster reactions and new synthesis and processing routes. We demonstrate the effects of glow (cold) and spark (hot) discharge milling on particle fracture for brittle, low-conductivity materials and ductile metals. Glow discharge milling was found to promote solid–gas reactions whereas spark discharge milling promotes fast fracturing, recrystallization, mineral reduction and solid–solid reactions.

A range of conventional and new milling and attriting devices has been developed for the initiation of various types of mechanically induced chemical reactions^{1,3,8}. These include planetary mills, where ball motion is generally chaotic, and vibrational mills, which may use either rods or balls. Milling performed under a controlled atmosphere can be used to initiate solid–solid, solid–liquid or solid–gas reactions; however, much work is still required in order to understand the complex processes occurring during such reactions. One common element of reactive and non-reactive milling is that, during the milling process, particles are generally deformed or broken up using some combination of shearing and impact.

For this work, we selected laboratory mills that produce distinct

milling modes under a controlled electrical discharge within a controlled atmosphere. One such device was a modified magneto-mill⁸ in which external magnets control the ball motion and restrict particle–ball interaction to some combination of impact and/or shearing (Fig. 1a). An advantage of using a mill of this type for discharge experiments is that the controlled ball movement makes easy the connection of electrical potential to the balls using wire slip-springs, which slide over the ball surfaces (Fig. 1b). A second device used in this work was a modified vibrational mill with a milling mode that mainly consisted of ball–particle impact under a vibrating rod with a hemispherical end. The power supply used in this study was connected through an alternating current (a.c.) high-voltage transformer, generating typically 30 kV, 50 Hz impulses within the microampere range. During milling, micro-gaps between moving and/or vibrating balls, and between balls and the milling chamber wall, occur, resulting in either an arc discharge or a discharge that was predominantly the glow discharge type (Fig. 1c).

During arc (or spark) discharge milling (above about 0.1 MPa gas pressure), the conduction path is through the balls, the ball–chamber gap and the powder particles to the mill chamber. Individual avalanche discharge channels result in rapid Joule heating. Subsequent particle fragmentation is believed to occur through some combination of vaporization or local melting associated with the Joule heating, and excessive thermal and tensile stresses resulting from the combined effects of milling and local temperature variations.

Milling of brittle, nonconducting ceramics in this way resulted in extremely rapid particle break-up, consistent with the fracturing process described above, whereas milling of conducting metal particles resulted in more complex effects. Figure 2 compares scanning electron micrographs (SEM) obtained from the electrical insulator alumina (a, c, e) with those obtained from nickel (b, d, f). In each case, the powders were milled for a short period in a non-reactive atmosphere, both with and without an electric discharge.

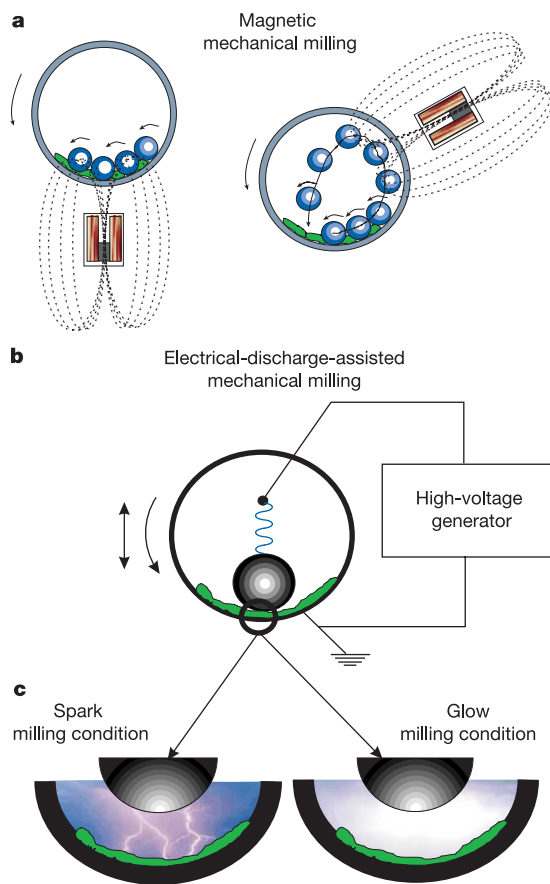


Figure 1 Schematic illustrations of magneto-ball milling and electric-discharge-assisted mills. **a**, Conventional milling under external magnetic field. External magnets control ball motion producing milling modes with various combinations of impact and shear. **b**, Electric-discharge-assisted mechanical milling method (either rotational or vibrational type mill). **c**, Type of milling is characterized by type of electric discharge, either hot milling under spark discharge or cold milling under glow discharge.

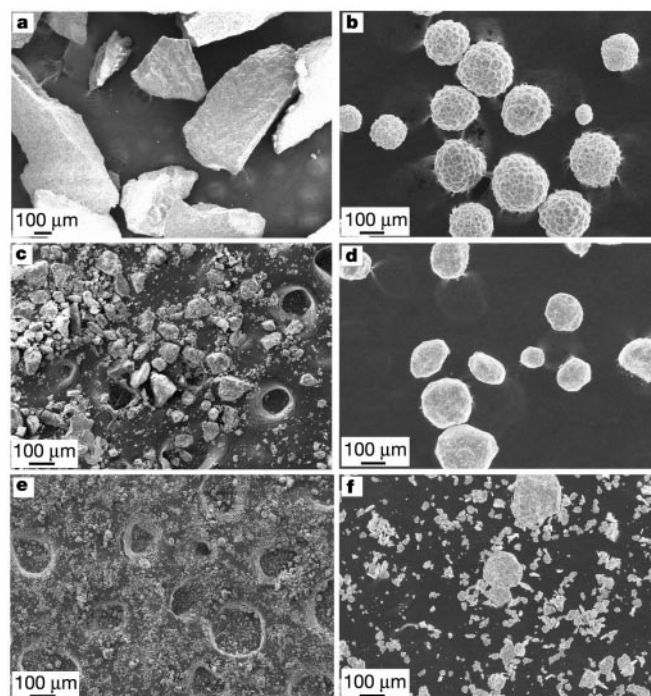


Figure 2 Fracturing resulting from spark discharge milling of low-electrical-conductivity particles and high-electrical-conductivity metal powders. **a**, **b**, SEMs of starting materials, alumina (**a**) and electrolytic nickel (**b**). **c**, **d**, Alumina milled for 1 min with no electric discharge (**c**) and nickel milled for 20 min without discharge (**d**). **e**, **f**, Alumina (**e**) and nickel (**f**) powders milled for 1–2 min with spark discharge.

Milling of Al_2O_3 for around 1 min without a discharge resulted in brittle fracture and a reduction in average powder particle size from around $300\text{ }\mu\text{m}$ to around $100\text{ }\mu\text{m}$ (Fig. 2c), while conventional milling of Ni for a short period resulted in particle deformation and smoothing of surfaces (Fig. 2d). Milling of Al_2O_3 with a discharge resulted in the extremely rapid break-up of particles and after 1 min most were within the size range $1\text{--}10\text{ }\mu\text{m}$ (Fig. 2e). The fracturing mechanisms operative during discharge milling of high-electrical-conductivity metals were more difficult to characterize. Discharges were found to induce surface roughening and surface fractures, resulting in chipping and shaving of metallic particles which were originally spherical in shape, apparently caused by fracture approximately concentric to the surface of particles (Fig. 2f). These effects were presumably associated with sparking in the vicinity of the outer regions of the particles.

It is well established that mechanical attrition or mechanical milling can generate solid–solid reactions, such as alloying or the direct formation of new phases. The morphology and structure of such reaction products can vary significantly and product types include: (1) microcrystalline powders, comprising small particles containing micrometre-sized crystals; (2) nanostructural powders, comprising micrometre- and submicrometre-size-particles containing a mixture of nanocrystals and disordered grain boundary phases; and (3) amorphous particles, with no long-range order. In this study, we found that discharge-assisted milling resulted in new reactions, which were not achieved during simple mechanical milling.

In one experiment, a pre-milled powder of Fe and B was used.

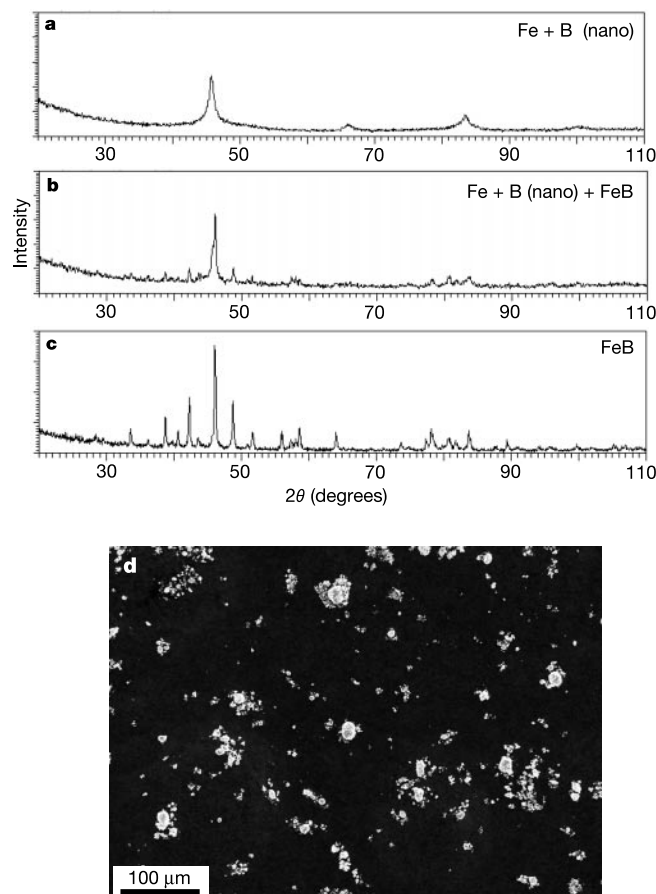


Figure 3 Mechanical milling of 50 at% Fe with 50 at% B. **a**, X-ray diffraction (XRD) pattern obtained from elemental powders of Fe + B magneto-milled for 200 h under impact. **b**, **c**, Fe + B magneto-milled for 200 h under impact and then spark milled for 15 min (**b**) and 30 min (**c**). **d**, SEM of sample shown in **c**.

Elemental powders were milled for 200 h in an argon atmosphere using a magneto-mill working in high-energy impact mode. Figure 3a shows an X-ray diffraction (XRD) pattern obtained from the end-milling product. After 200 h milling the Fe peaks are very broad. This type of pattern is consistent with the formation of nanostructural powder containing nanocrystalline Fe and B phases. Prolonged milling for up to 500 h did not alter this pattern, indicating that milling of Fe and B does not result in a new alloying reaction. However, if at this stage the high voltage is switched on, after 15 min clearly visible X-ray peaks corresponding to the formation of the compound, FeB, appear, together with the remaining Fe peaks (Fig. 3b). After 30 min of discharge-assisted milling

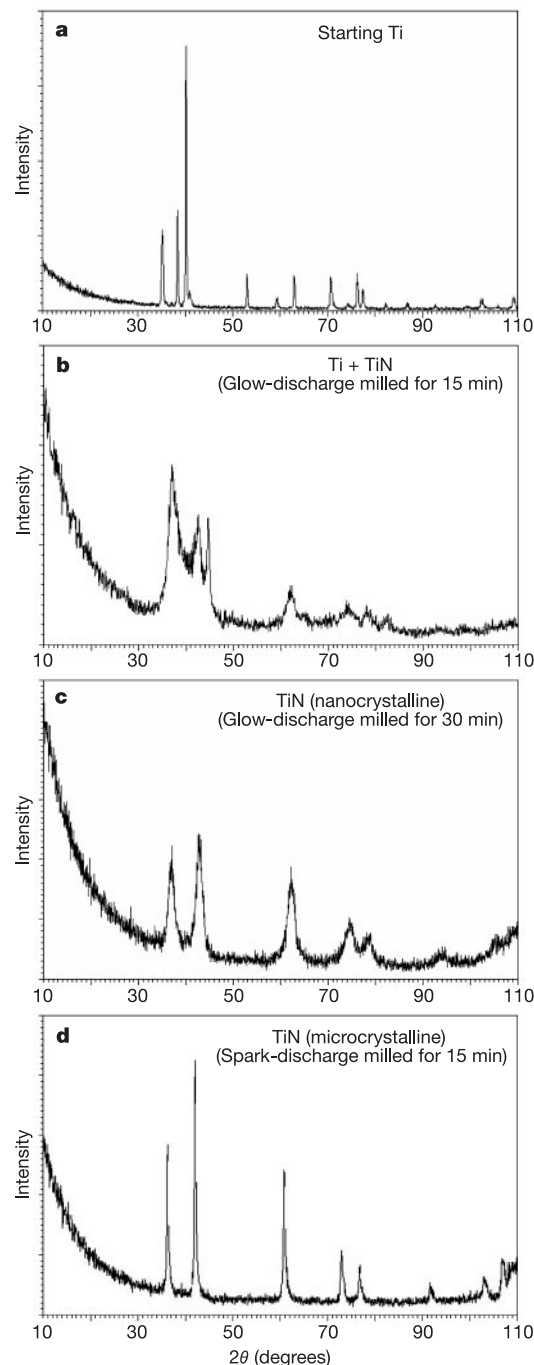


Figure 4 Solid–gas reaction paths during discharge-assisted milling of Ti in nitrogen. XRD patterns were obtained from starting powder (**a**), Ti glow-discharge milled for 15 min (**b**), Ti glow-discharge milled for 30 min (**c**), and Ti glow-discharge milled for 30 min and then spark milled for 15 min (**d**).

only the set of X-ray peaks corresponding to the phase FeB is visible (Fig. 3c). Figure 3d is an SEM of the sample after spark-assisted milling for 30 min, showing the powder product morphology.

A wide range of other solid–solid reactions could be induced by electric-discharge milling, for example, the rapid formation of NiSi from elemental Ni and Si powders (this reaction is not achievable via low-energy mechanical milling) and the direct transformation of carbon into a range of crystallized forms. In the carbon experiments, the milling products included a small fraction of unusual nano-fragments, including nanotubes, of a type apparently containing a high density of stacking faults perpendicular to the long axes of the tubes. We are currently investigating ways of increasing the yield of unusual forms of carbon which may be produced by electric-discharge milling.

Several workers have described the formation of metal nitrides during milling of metals in nitriding gas environments^{7,9,10}. Irrespective of the milling method, only a limited number of reactive metals, including Ti, Zr and V, have been successfully nitrided by prolonged milling of elemental powders in nitrogen. During milling, these metals promote catalytic dissociation of nitrogen molecules into monatomic nitrogen on the fractured metal surface. Nitriding of other metals by ball milling in nitrogen is generally more difficult. Materials such as Si and Fe can be nitrated more rapidly when the nitrogen atmosphere is replaced with ammonia^{9,10}. In Fe and Si, local rises in temperature associated with the milling process are believed to help induce decomposition of ammonia into monatomic nitrogen, which reacts with newly created surfaces of milled material. However, a side effect of this reaction is the formation of hydrogen, which reacts not only with the milled material but also with the milling cylinder surface (usually steel or stainless steel). This results in embrittlement of the mill chamber and subsequent chipping, and also increases iron contamination

of the powder product.

In the case of discharge-assisted milling, nitriding processes appear quite different. Under spark-discharge conditions a relatively hot plasma exists along the spark path. This is believed to result in the formation of monatomic nitrogen if the milling process is carried out in nitrogen, or decomposition into monatomic hydrogen and nitrogen for milling in ammonia gas. Characteristics of spark milling include the condition that the reaction area around the spark be quite small, so powder particles may be heated to very high temperatures and presumably nitrided in small localized regions. In contrast, the temperature generated around a glow discharge is relatively low and the plasma itself is in contact with a larger area of particle surfaces where possible reactions may be initiated.

To investigate the influence of discharge milling on nitriding, Ti powder was milled under both electric-spark and glow-discharge conditions. After 2 h using a spark discharge no evidence of nitriding was observed. The associated XRD pattern showed only Ti peaks, with a slight broadening in comparison with peaks associated with non-milled Ti powder (Fig. 4a). This indicates that the Ti powder did not react significantly with nitrogen. In contrast, when the pressure was reduced and a discharge mainly of the glow type was induced, peaks associated with formation of TiN phase appeared after 15 min (Fig. 4b), and after 30 min the XRD pattern showed only broad peaks corresponding to the nanostructural form of TiN phase (Fig. 4c), with crystallite size around 30 nm. If at this stage of milling the spark (hot) milling condition is switched on, recrystallization occurs and after 15 min of milling X-ray peaks become sharp, indicating transformation from the nano- to microcrystalline phase (Fig. 4d). Similar results were obtained for Si powder, nanocrystalline Si₃N₄ being formed by milling of Si in nitrogen gas under glow-discharge conditions, and a

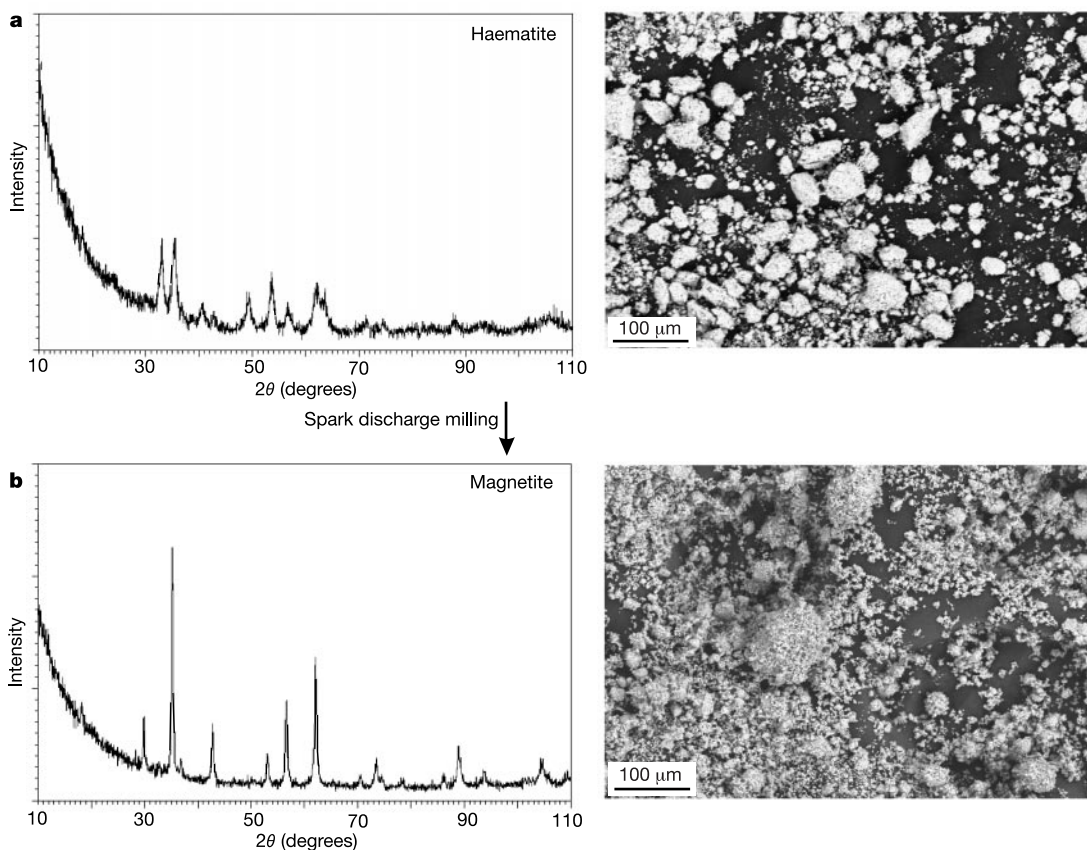


Figure 5 Reduction reaction of haematite to magnetite caused by spark discharge milling. **a**, XRD pattern and SEM from magnetite conventionally milled for 20 h; **b**, the same sample as in **a**, spark-discharge milled for 30 min in nitrogen.

fine-grained microcrystalline Si_3N_4 product produced after an additional short period of spark-discharge milling. Silicon can be nitrided in ammonia by conventional magneto-ball milling under low-energy shearing conditions to form an amorphous-like product¹⁰. The powder can then be hot-pressed to form monolithic microcrystalline Si_3N_4 . However, the more direct route of hot-pressing microcrystalline Si_3N_4 that is produced by discharge milling may be more attractive because such Si_3N_4 is less likely to oxidize during processing and because there is no requirement to remove hydrogen.

Spark-discharge milling was successfully employed for initiation of reduction processes. Results obtained for haematite are shown in Fig. 5. We used pre-milled powder, magneto-milled under conventional impact mode in air. After 20 h, XRD indicated broad peaks corresponding to nanostructural haematite. This powder was subsequently milled for 30 min in nitrogen under spark milling conditions, resulting in formation of magnetite or maghaematite. As indicated by the associated SEMs, the corresponding change in particle morphology is from clumps of nanocrystalline powder (Fig. 5a) to finer particles containing coarser, submicrometre crystallites (Fig. 5b).

There are promising applications of this electrical discharge method for materials and minerals processing. Handling of ultra-fine ball milled powders with large surface areas is generally difficult owing to their high susceptibility to structural changes, particularly oxidation, even in inert atmospheres containing only trace impurities such as water vapour or oxygen. Such problems may potentially be overcome if discharge-assisted milling can be applied for the direct formation of more stable phases. A general method might involve conventional milling to produce a reactive ultrafine or nanostructural product followed by final milling under an electrical discharge to generate particular heat-induced transformations, such as recrystallization, grain growth, relaxation of internal stresses or the direct formation of new phases by solid–solid reaction. Such a technique may avoid problems associated with handling and subsequent treatment of highly reactive powders. Areas of minerals processing which might benefit from scaled-up versions of discharge-assisted milling range from more energy-efficient methods of ore fracturing to alternative routes of minerals reduction. □

Received 29 September 2001; accepted 16 July 2002; doi:10.1038/nature00985.

1. Zoz, H. Attritor technology—latest developments. *Mater. Sci. Forum* **179–181**, 419–424 (1995).
2. Koch, C. C. Materials synthesis by mechanical alloying. *Annu. Rev. Mater. Sci.* **18**, 121–143 (1989).
3. Basset, D., Matteazzi, P. & Mani, F. Designing a high energy ball-mill for synthesis nanophase materials in large quantities. *Mater. Sci. Eng. A* **A168**, 149–152 (1993).
4. Pelletier, J. et al. New trends in DECR plasma technology: applications to novel duplex treatments and process combinations with extreme plasma specifications. *Surf. Coat. Technol.* **139**, 222–232 (2001).
5. Takaki, K., Taguchi, D. & Fujiwara, T. Voltage–current characteristics of high-current glow discharges. *Appl. Phys. Lett.* **78**, 2646–2648 (2001).
6. Mishra, R. S. & Mukherjee, A. K. Electric pulse assisted rapid consolidation of ultrafine grained alumina matrix composites. *Mater. Sci. Eng. A* **A287**, 178–182 (2000).
7. El-Eskandarany, M. S., Sumijama, K., Aoki, K. & Suzuki, K. Syntheses of full-density nanocrystalline titanium nitride compacts by plasma-activated sintering of mechanically reacted powder. *Met. Mater. Trans. A* **29A**, 1973–1981 (1998).
8. Calka, A. & Radlinski, A. P. Universal high performance ball milling device and its application for mechanical alloying. *Mater. Sci. Eng. A* **A134**, 1350–1353 (1991).
9. Wexler, D., Calka, A. & Colburn, S. J. Rapid sintering of nanostructural powder to form Si_3N_4 . *Mater. Sci. Forum* **269–272**, 219–224 (1998).
10. Li, Z. L., Williams, J. S. & Calka, A. The role of hydrogen and iron in silicon nitridation by ball milling. *J. Appl. Phys.* **81**, 8028–8034 (1997).

Acknowledgements

We acknowledge R. deJong, R. Kinnell and S. Selby for their contributions towards construction of the discharge milling devices. We are especially grateful to D. Dunne for his support and leadership. This project was supported by funding from the Australian Research Council.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.C. (e-mail: aoc2107@uow.edu.au).

Acceleration of rain initiation by cloud turbulence

G. Falkovich*, A. Fouxon* & M. G. Stepanov*†

* Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

† Institute of Automation and Electrometry, Novosibirsk 630090, Russia

Vapour condensation in cloud cores produces small droplets that are close to one another in size. Droplets are believed to grow to raindrop size by coalescence due to collision^{1,2}. Air turbulence is thought to be the main cause for collisions of similar-sized droplets exceeding radii of a few micrometres, and therefore rain prediction requires a quantitative description of droplet collision in turbulence^{1–5}. Turbulent vortices act as small centrifuges that spin heavy droplets out, creating concentration inhomogeneities^{6–14} and jets of droplets, both of which increase the mean collision rate. Here we derive a formula for the collision rate of small heavy particles in a turbulent flow, using a recently developed formalism for tracing random trajectories^{15,16}. We describe an enhancement of inertial effects by turbulence intermittency and an interplay between turbulence and gravity that determines the collision rate. We present a new mechanism, the ‘sling effect’, for collisions due to jets of droplets that become detached from the air flow. We conclude that air turbulence can substantially accelerate the appearance of large droplets that trigger rain.

The local distribution of droplets over sizes, $n(a, t, \mathbf{r}) = n(a)$, changes with condensation and coalescence according to refs 1, 2 (see Table 1 for definitions of variables):

$$\frac{\partial n(a)}{\partial t} = -q \frac{\partial n(a)}{\partial a} \frac{1}{a} - \frac{\partial}{\partial \mathbf{r}} \mathbf{v} n + \int da' \left[\frac{K(a', a'') n(a') n(a'')}{2(a''/a)^2} - K(a', a) n(a') n(a) \right] \quad (1)$$

Here, $\mathbf{v}(t, \mathbf{r})$ is the droplet velocity at point $\mathbf{r}$ at time t , q is proportional to the supersaturation and the diffusivity of the vapour and $a'' = (a^3 - a'^3)^{1/3}$. The collection kernel is proportional to the collision kernel, which is the product of the target area and the relative velocity Δv of droplets before the contact: $K(a_1, a_2) \approx \pi(a_1 + a_2)^2 \Delta v$. For droplets larger than couple of micrometres across, brownian motion can be neglected and the collision kernel in still air is due to gravitational settling^{1,2}: $K_g(a_1, a_2) = \pi(a_1 + a_2)^2 E(a_1, a_2) |u_g(a_1) - u_g(a_2)|$. When the Reynolds number of the flow around the droplet, $Re_a \equiv u_g a / \nu$, is not too large and concentration is small enough ($na^3 Re_a^{-2} \ll 1$) the settling velocity is due to the balance of gravity and friction: $u_g = g\tau$ with the Stokes time $\tau = (2/9)(\rho_0/\rho)(a^2/\nu)$. Here ρ_0 , ρ are water and air densities respectively. Hydrodynamic interaction between approaching droplets is accounted for in K_g by the collision efficiency E , for which values can be found in refs 1, 17.

Cloud condensation nuclei are typically of micrometre or sub-micrometre size and the initial stage of droplet growth is solely due to condensation: $n(a, t) = af(a^2 - 2qt)$ with the function f determined by the initial distribution. An important conclusion is that while the distribution shifts to larger sizes, it keeps its small width over a^2 (a few micrometres squared or less). It would take many hours for condensation to grow millimetre-size raindrops, especially with the account of vapour depletion. Such growth is supposed to come from coalescence but because $K_g \propto |a_1^2 - a_2^2|$ the gravitational collision rate is strongly suppressed for droplets that are close in size. With narrow local size distributions in cloud cores, rain would not start in still air for many hours. There is then the

long-standing problem of the bottleneck in the transition from condensation to coalescence stage^{1–5,8–11,18–20} which we discuss here.

In some cases, droplets with substantially different sizes may appear, owing to the existence of ultra-giant nuclei^{18,19}. Another possibility that we consider here is that turbulence-induced collisions^{8–11,20} may occur. Both velocity and concentration of droplets fluctuate in a random flow. To provide meteorology with an effective computational tool, theoretical physics is expected to produce the condensation–coagulation equation (1) averaged over space. Here we derive analytically the averaged equation—that is, we obtain the effective collision kernel $\bar{K} = \langle K(a_1, a_2) n_1 n_2 \rangle / \langle n_1 \rangle \langle n_2 \rangle$ —and solve it numerically to demonstrate the changes in the average distribution $n(a, t)$ brought about by turbulence.

For the basic discussion of cloud turbulence see the reviews in refs 3–5 and the references therein. Turbulence intensity can be characterized by the energy input rate ε which determines the root-mean-square (r.m.s.) velocity gradient: $\lambda \approx (\varepsilon/\nu)^{1/2}$. We consider small droplets with the Stokes number $St = \lambda\tau$ (which characterizes mean droplet inertia) smaller than unity. If inertia is neglected, droplets follow the incompressible air flow and their concentration is uniform. Droplet motion in the air flow gradient s then provides $\Delta v \approx s(a_1 + a_2)$ and gives the mean collision kernel²⁰ $\langle K_i \rangle \approx \lambda(a_1 + a_2)^3$. Inertia deviates droplets from the air flow, adding a contribution to Δv proportional to τs the mean value of which is St , that is, small²⁰. Hence Saffman and Turner²⁰ concluded that only extremely energetic turbulence with $\varepsilon > 2,000 \text{ cm}^2 \text{ s}^{-3}$ would produce a noticeable collision kernel. We note, however, that it is $\bar{K}$ rather than $\langle K \rangle$ that determines the mean collision rate. Inertial deviation of droplets from the air flow leads to fluctuations in droplet concentration, characterized by the factor $k_{12} = \langle n_1 n_2 \rangle / \langle n_1 \rangle \langle n_2 \rangle > 1$, which may be large. Concentration fluctuations have been observed in experiments and numerical simulations^{5–12} and described analytically for same-size droplets in low Reynolds flow without gravity^{13,14}: $k(a, a) = \langle n^2 \rangle / \langle n \rangle^2 = (\eta/a)^{St^2}$, where $\eta \approx (\nu^3/\varepsilon)^{1/4}$ is the mean correlation scale of velocity gradients.

The Reynolds number of cloud turbulence is $Re = uL/\nu$, where u is air velocity and L is the outer scale comparable to the cloud size. In the atmosphere, Re is large (10^6 – 10^8), that is, turbulence is intermittent and the statistics is very non-gaussian, with a substantial probability of gradients far exceeding λ . The role of gravity can be characterized by the ratio of the small-scale turbulent velocity to the settling velocity^{12,21} $\epsilon \equiv (\varepsilon\nu)^{1/4}/u_g$; this parameter can be both larger and smaller than unity for $a = 1$ – $100 \mu\text{m}$ and $\lambda = 1$ – 20 s^{-1} .

Here we derive the factor k_{12} for droplets under gravity in high- Re flow. We show that contribution of large gradients can significantly increase k_{12} compared with the low- Re case and that gravity provides for a sharp maximum of k_{12} at $a_1 = a_2$. We also describe a new inertial mechanism of collisions due to rare events with large gradients ($s \approx \tau^{-1} \gg \lambda$) that produce jets of droplets initially accelerated by the air flow and then detached from it. That gives an additive contribution K_i into $\bar{K}$. We call this the ‘sling effect’ and show that turbulence intermittency can make K_i substantial even at small St . No realistic direct numerical simulations are possible for

droplets in high- Re turbulence, so analytical derivations are indispensable. We derive $\bar{K}(\epsilon, St, Re)$ and show that the turbulence-induced collision rate can be substantial even for small droplets in moderate turbulence when St is small.

The field $\mathbf{v}(\mathbf{r}, t)$ giving the velocity of a droplet located at $\mathbf{r}$ at time t satisfies the equation^{6,22} $\partial_t \mathbf{v} + (\mathbf{v} \cdot \nabla) \mathbf{v} = (\mathbf{u} - \mathbf{v})/\tau + g\mathbf{z}$, where $\mathbf{z}$ is a unit vector pointing downwards. The gradients $\sigma = \nabla \mathbf{v}$ and $s = \nabla v$ taken at a droplet’s trajectory are related as follows: $\dot{\sigma} + \sigma^2 = (s - \sigma)/\tau$. When $|\sigma| \ll \tau^{-1}$, it has a smooth evolution determined by $\sigma(t) = \int^t dt' \exp[(t' - t)/\tau] s(t')/\tau$. If, however, $|\sigma| > \tau^{-1}$ then the inertial term σ^2 dominates and may lead to an explosive evolution $\sigma(t) \propto (t_0 - t)^{-1}$ that produces shock in $\mathbf{v}(\mathbf{r})$ and singularities in $n(\mathbf{r})$.

The probability P of an explosive event is that of large and persistent gradients s . The correlation time $\tau_c(s)$ of the air flow gradient is given by the minimum between the turnover time $|s|^{-1}$ and the time $l(s)/u_g$ needed for droplets to cross the region $l(s) \approx \sqrt{\nu/|s|}$ over which s is correlated. The gradient s that leads to $|\sigma| \approx \tau^{-1}$ must either exceed the threshold described by the extrapolation formula $s_b = [\tau^{-2} + \lambda^2 \epsilon^{-4}]^{1/2}$ or be larger than $1/\tau$ and occupy the region in space $l(s_b)s_b/s$. Because the only available data are on the single-point probability density function (PDF) $\mathcal{P}(s)$ we estimate the probability of explosion from below: $P \equiv \int_{1/\tau}^{\infty} \mathcal{P}(|\sigma|) d|\sigma| \approx \tau s_b \int_{s_b} \mathcal{P}(|s|) d|s|$, where the prefactor τs_b appears because $s > s_b$ can occur at any moment within the interval τ . Once a fluctuation with a negative eigenvalue $\sigma_i < -\tau^{-1}$ occurs, the inertial term σ^2 exceeds the driving term s_i/τ and the friction term $-\sigma_i/\tau$, which corresponds to a free motion of droplets along the direction of σ_i on a timescale of order τ . A negative velocity gradient means that faster droplets catch up with slower ones, creating a cubic singularity²³ in the relation between the current coordinate $x(t)$ and the initial one y : $x = y^3/3l^2 - yt/\tau$. Here $l = l(s_b)$ is the correlation length of $|\sigma| = 1/\tau$ and t is counted from the moment of singularity. Using $n(x, t) = n(y) |\partial y / \partial x|$ and $|\sigma| = |(\partial \mathbf{v} / \partial y)(\partial y / \partial x)| \approx \tau^{-1} |\partial y / \partial x|$ we find the contribution of the preshocks ($t < 0$) into the collision rate: $\langle |\sigma| n^2 \rangle \approx P \tau^{-1} \int dx \int_{-\infty}^0 dt (\partial y / \partial x)^3 \langle n^2(y) \rangle / l \tau = P \tau^{-1} (l/a)^{1/3} \langle \tilde{n}^2 [a(l/a)^{2/3}] \rangle$. Formula $\langle |\sigma| n^2 \rangle$ assumes smoothness over the scale a , so we introduced $\tilde{n}[\delta y]$ coarse-grained over δy , taking $\delta y \approx a(l/a)^{2/3}$, which corresponds to $\delta x \approx a$.

After the shock ($t > 0$), folds appear in the map $y(x)$ in the region $|x| < 2l(t/\tau)^{3/2}/3$: for every x value there are three y values which correspond to the three groups of droplets that came from different places and have different velocities. Nearby droplets from the same group have $\Delta v \approx \sigma a$ and contribute $\langle |\sigma| n^2 \rangle \approx P \tau^{-1} (l/a)^{1/2} \langle \tilde{n}^2 [(la)^{1/2}] \rangle$. However, this contribution and that of preshocks are both less than that given by collisions of droplets from different groups. Groups coming from afar appear because droplets

Table 1 Parameters

Quantity	Units	Description
$n(a)$	cm^{-4}	Distribution of droplets over sizes in a unit volume
a	cm	Droplet radius
t	s	Time
$\mathbf{r}$	cm	Spatial coordinate
q	$\text{cm}^2 \text{ s}^{-1}$	Rate of condensational growth
K	$\text{cm}^3 \text{ s}^{-1}$	Collision kernel
u_g	cm s^{-1}	Terminal fall velocity
g	cm s^{-2}	Acceleration of gravity
ν	$\text{cm}^2 \text{ s}^{-1}$	Air viscosity
ε	$\text{cm}^2 \text{ s}^{-3}$	Energy dissipation rate per unit mass
$\mathbf{u}(\mathbf{r}, t)$	cm s^{-1}	Air velocity

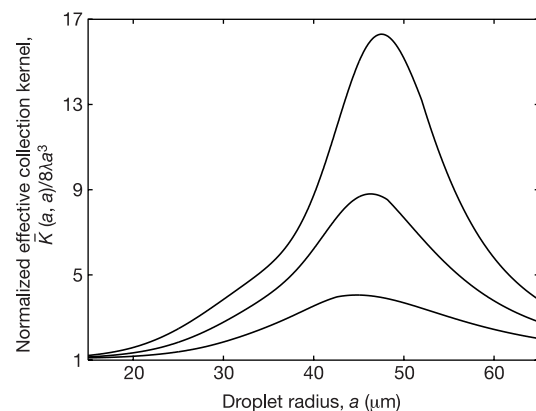


Figure 1 Normalized effective collection kernel for equal-size droplets at $Re \approx 10^6$ according to equations (2), (3), (4) and (6). From bottom to top, $\lambda = 10, 15$ and 20 s^{-1} .

are shot out of curved streamlines with too high a centrifugal acceleration, an effect known to anyone who has used a sling to throw stones. That is why we call this the ‘sling effect’. Droplets separated at the beginning of the free motion by a distance $\approx l$ have $\Delta v \approx l/\tau$ and provide for the inertial collision kernel:

$$K_i(a_1, a_2) \approx \pi(a_1 + a_2)^2 E' (P_1 l_1 / \tau_2 + P_2 l_2 / \tau_1) / 2. \quad (2)$$

Subscripts denote different droplet sizes, $\tau(a_1) = \tau_1$, $P_1 = P(|\sigma| > 1/\tau_1)$. Because the velocities and the concentrations of the different groups are uncorrelated, $\langle K_i(a_1, a_2) n_1 n_2 \rangle = K_i(a_1, a_2) \langle n_1 \rangle \langle n_2 \rangle$. The collision efficiency E' in equation (2) can be expressed via E taken for the effective sizes, giving the same Δv . We thus obtain $E' \approx 0.93$ – 0.98 in the interval 15 – $100 \mu\text{m}$ for collinear velocities (non-collinearity further increases E' ; ref. 3), so with our accuracy, $E' \approx 1$. We note that $K_i(a, a)$ is larger than the contribution due to droplets from the same group by the factor $\sqrt{l/a} \langle \tilde{n}^2[l] \rangle / \langle \tilde{n}^2[(la)^{1/2}] \rangle$. It is shown below that $\langle \tilde{n}^2[r] \rangle \propto r^{-\alpha}$ with $\alpha < 1$ so that K_i indeed dominates. The ratio $K_i/\lambda a^3 \approx P \text{St}^{-1}(l/a)$ has the smallness of P compensated by two large factors and can exceed unity even at small St . Most importantly, $K_i(a, a) \neq 0$.

We now describe concentration fluctuations and derive k_{12} . Because of inertia, the divergence of $\mathbf{v}$ is nonvanishing⁶, albeit small: $\text{tr}\sigma = -\int \exp[(t' - t)/\tau] \text{tr}\sigma^2(t') dt' \ll |\sigma|$ at $|\sigma| \ll 1$. Negative $\text{tr}\sigma^2$ corresponds to elliptic flows (vortices) which act as centrifuges decreasing n . Droplets concentrate in hyperbolic regions (between the vortices) where $\text{tr}\sigma^2 > 0 > \text{tr}\sigma$. Clusters of droplets are created with sizes not exceeding η , as follows from theory^{13,14} and as seen in numerics²⁴ and observations²⁵. We note in passing that as clusters do not exceed η the fluctuations of droplet concentration do not produce significant fluctuations in vapour concentration (because the vapour diffusivity is comparable to the air viscosity) so that droplet distribution over sizes cannot be significantly broadened during the condensation stage; see also ref. 26. Clustering can be readily understood: a compressible flow with lagrangian chaos creates a fractal concentration, the so-called Sinai–Ruelle–Bowen measure¹⁶. The moments of the fractal measure behave as powers of the scale ratio: $\langle \tilde{n}[r]^\beta \rangle \approx \langle n \rangle^\beta (\eta/r)^{\gamma(\beta)}$, where $\gamma(\beta)$ is convex and $\gamma(0) = \gamma(1) = 0$. As $\gamma'(0)$ is negative¹³ (it is proportional to the sum of the backward-in-time Lyapunov exponents of the $\mathbf{v}$ -flow), then $\alpha \equiv \gamma(2) > 0$. Droplets of different sizes have additional relative velocity $|\tau_1 - \tau_2|(g + \lambda^2 \eta)$ that stops clustering at $r \approx |\tau_1 - \tau_2|(g + \lambda^2 \eta)/\lambda_d$. We thus find:

$$k_{12} \approx \langle \tilde{n}_1(a) \tilde{n}_2(a) \rangle / (\langle n_1 \rangle \langle n_2 \rangle) \approx [a/\eta + (\lambda + g/\lambda \eta)|\tau_1 - \tau_2| \epsilon^{-1}]^{-\alpha} \quad (3)$$

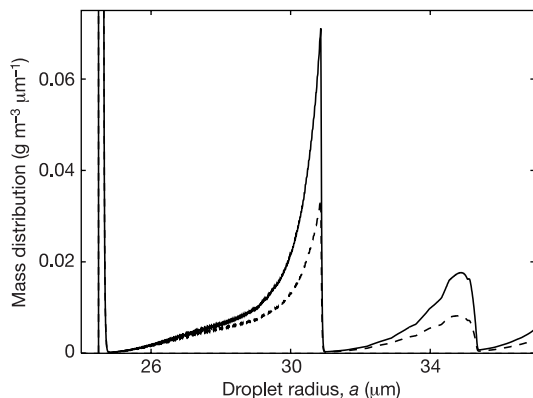


Figure 2 Distribution over sizes after 10 min. The dashed line is the solution of equation (1) with the mean-field collision kernel and the solid line is the solution of equation (1) with equation (4).

We distinguish a_1 from a_2 only in $|\tau_1 - \tau_2|$ giving the sharpest dependence. The exponent α is described by equation (6), derived in the Methods section below. For sufficiently small droplets ($\text{St} < 1$ and $\epsilon > 1$) and not very high Re , we have $\alpha \approx \text{St}^2 F_3$, where $F_3 \approx \lambda^{-3} \int |\mathbf{s}|^3 \mathcal{P}(\mathbf{s}) d\mathbf{s}$ is a growing function of Re that describes how turbulence intermittency amplifies the effect of small droplet inertia. At low Re and $\text{St} < 1$, α does not depend on ϵ , so the only dependence $k_{11}(\epsilon)$ can come from shock contribution and has to be weak (logarithmic), which agrees with numerics¹². At large Re , both α and k_{12} have a maximum at $\text{St} \approx \epsilon^2$.

We now write the effective collision kernel for small heavy particles in turbulence:

$$\bar{K}(a_1, a_2) \approx k_{12} \lambda (a_1 + a_2)^3 + k_{12} K_g + K_i \quad (4)$$

To compare with numerics done for low Re without gravity^{9,10} we use equations (2), (3), (4) and (6) with $P \approx \exp(-\text{St}^2)$ and $l \approx \eta \text{St}^{1/2}$. Analytics and numerics agree well, showing fast growth of $\bar{K}$ with St at $\text{St} < 1$ and a (broad) maximum at $a_1 = a_2$ (refs 9, 10). As St approaches unity, $k_{12} \gg 1$ and $K_i \gg \lambda(a_1 + a_2)^3$, which explains the observation⁹ that contributions of both preferential concentration and relative velocity are important. Gravity suppresses K_i increasing s_b at $\epsilon^2 < \text{St}$. It also makes k_{12} a sharp function of $|a_1 - a_2|$ so that the gravitational collision rate is only weakly enhanced by preferential concentration, because $k_{12} > 1$ where $K_g \approx 0$. At small St , we can also neglect the turbulence-induced increase of the vertical flux¹².

For practical applications to high- Re flows, α and K_i have to be evaluated with $\mathcal{P}(\mathbf{s})$ determined experimentally. To make an estimate from below, we numerically evaluate equations (2), (3), (4) and (6) for moderate turbulence with $\text{Re} \approx 10^6$, taking $\mathcal{P}(\mathbf{s})$ from ref. 27. Figure 1 shows the effective collision kernel $\bar{K}$ normalized by $8\lambda a^3$. The normalized kernel has a maximum at $\text{St} \approx \epsilon^2$ which corresponds to the balance between inertia and gravity when the Stokes time τ is the universal value $(\nu/g^2)^{1/3}$. Droplets of such size take time τ to fall through the vortex with a turnover time τ . The effect of centrifugal force is less both for smaller droplets (which are less inertial) and for larger ones (which spend less time inside the vortex). This is to be contrasted with the maximum at $\text{St} \approx 1$ in low-Reynolds numerics without gravity^{9,10}. How inertia and gravity influence droplet settling is discussed in refs 12, 21, 28; see also refs 29, 30 on the role of turbulence.

We see that the interplay between gravity and turbulence intermittency makes inertial enhancement of the turbulence-induced collision rate significant only in the restricted interval of droplet sizes that depends on the air density (between 20 and $60 \mu\text{m}$ for $\rho = 10^{-3} \text{g cm}^{-3}$). The condensation–coagulation bottleneck is expected precisely in this interval, so turbulence must be able to alleviate it. To illustrate the effect, we solve space-averaged equation (1) numerically with the mean-field collection kernels $K_g + \langle K_i \rangle$ (dashed line in Fig. 2) and with $\bar{K}$ (solid line in Fig. 2) for $\lambda = 20 \text{ s}^{-1}$, $\eta = 6 \text{ mm}$ and $q = 5 \times 10^{-9} \text{ cm}^2 \text{ s}^{-1}$. We took 50 droplets per cm^3 with initial sizes in the interval 2 – $3 \mu\text{m}$. Even for such relatively low level of turbulence and small size of droplets, the difference in coalescence-produced secondary peaks is apparent after only 10 min. The main peak is at $a = 25 \mu\text{m}$. The number density of coalescence-produced droplets is 1.06 cm^{-3} with the newly found $\bar{K}$ versus 0.64 cm^{-3} with the old mean-field values.

We thus conclude that turbulence-induced inertial effects can substantially accelerate the transition from condensation to coalescence stage in the interval of few tens of micrometres. Our results are valid for a low concentration of small droplets and not very energetic turbulence, conditions compatible with the data for most clouds^{1,2,4,5}, so we believe that equations (2), (3), (4) and (6) provide for an effective tool in rain prediction. □

Methods

To determine α from equation (3) we consider¹³ the prehistory of a cluster with the

smallest size r . It is formed by a gradual deformation of an η -size region with an initially uniform concentration. The shock contribution to the moment, $\langle \bar{n}^2[r] \rangle \propto \int dx (\partial y / \partial x)^2 \propto \ln(l/r)^{1/2}$, contains a logarithm that is of order unity in our case. Therefore, we neglect shocks and consider fluctuations with $|\sigma| < \tau^{-1}$. The smallest size of the region evolves as $\eta \exp[\lambda_d t]$, where λ_d is the most negative Lyapunov exponent estimated as $|\lambda_d| \approx \int dt \langle \tau \sigma^T(0) \sigma(t) \rangle \approx \lambda \epsilon = \lambda \epsilon (1 + \epsilon^2)^{-1/2}$. Therefore, concentration fluctuations accumulate during the time $\ln(\eta/r)/|\lambda_d|$. Because $\dot{n} = -n\tau\sigma$ in the droplets' frame and the contribution of each cluster to the spatial average is proportional to its volume $\exp[\int_0^t \tau \sigma(t') dt']$, we obtain:

$$\begin{aligned} \langle \bar{n}^2[r] \rangle &= \langle n \rangle^2 \left\langle \exp \left[- \int_0^{\ln(\eta/r)/\lambda_d} \tau \sigma(t') dt' \right] \right\rangle \\ &= \langle n \rangle^2 \left\langle \exp \left[\tau \int_0^{\ln(\eta/r)/\lambda_d} \tau \sigma^2(t') dt' \right] \right\rangle = \langle n \rangle^2 \left(\frac{\eta}{r} \right)^\alpha, \text{ for } \alpha \\ &\approx \tau^2 / |\lambda_d| \int dt \langle \tau \sigma^2(0) \tau \sigma^2(t) \rangle \end{aligned} \quad (5)$$

where we assumed that $\ln(\eta/r)/|\lambda_d|$ is much larger than the correlation time of $\tau \sigma^2$. The higher terms of the cumulant expansion cannot be parametrically larger than the estimate (5) since they contain integrals estimated as $\langle \tau \tau \sigma^2 [\tau \tau \sigma(t) \tau \sigma^2]^{2m+1} \rangle$, for $m \geq 1$, and both the correlation time $\tau_c(\sigma)$ and τ are less than $|\sigma|^{-1}$ in the integration domain. Moreover, if $\langle \tau^2 \tau_c(\sigma) (\tau \sigma^2)^2 \rangle$ is determined by $|\sigma| \ll \tau^{-1}$ then equation (5) is correct not only parametrically but also numerically. To evaluate α we express it via the single-time PDF $P(|\sigma|)$:

$$\alpha \approx (\tau^2 / \lambda \epsilon) \int_{|\sigma| < 1} d\sigma \sigma^4 \tau_c(\sigma) P(|\sigma|) \quad (6)$$

To relate $P(\sigma)$ to $\mathcal{P}(s)$ measured experimentally we note that $\tau_c(s) > \tau$ for $s < s^* = \tau^{-1} \min\{1, \epsilon^2/2\}$ so that $P(|\sigma|) = P(|s| = |\sigma|)$ there. At $s > s^*$ the fluctuations of s contributing to $P(\sigma)$ have $\tau_c(s) < \tau$ and can occur at any moment within τ . The extrapolation formulas at $St < \epsilon$ are $P(|\sigma|) = (1 + \sigma^2/s_*^2) \mathcal{P}(|s| = |\sigma| + \sigma^2/s_*)$ and $\tau_c(|\sigma|) = \tau + (|\sigma| + \lambda^{1/2} |\sigma|^{1/2} \epsilon^{-1})^{-1}$. Our theory is valid as long as $St \min\{1, \epsilon\} < 1$.

Received 28 February; accepted 11 July 2002; doi:10.1038/nature00983.

- Pruppacher, H. & Klett, J. *Microphysics of Clouds and Precipitation* (Kluwer, Dordrecht, 1998).
- Seinfeld, J. & Pandis, S. *Atmospheric Chemistry and Physics* (Wiley, New York, 1998).
- Pinsky, M., Khain, A. & Shapiro, M. Stochastic effects of cloud droplet hydrodynamic interaction in a turbulent flow. *Atmos. Res.* **53**, 131–169 (2000).
- Jonas, P. Turbulence and cloud microphysics. *Atmos. Res.* **40**, 283–306 (1996).
- Vaillancourt, P. A. & Yau, M. K. Review of particle-turbulence interactions and consequences for cloud physics. *Bull. Am. Met. Soc.* **81**, 285–298 (2000).
- Maxey, M. R. The gravitational settling of aerosol particles in homogeneous turbulence and random flow field. *J. Fluid Mech.* **174**, 441–465 (1987).
- Squires, K. & Eaton, J. Measurements of particle dispersion from direct numerical simulations of isotropic turbulence. *J. Fluid Mech.* **226**, 1–35 (1991).
- Pinsky, M. & Khain, A. Turbulence effects on droplet growth and size distribution in clouds—a review. *J. Aerosol Sci.* **28**, 1177–1214 (1997).
- Sundaram, S. & Collins, L. Collision statistics in an isotropic particle-laden turbulent suspension. *J. Fluid Mech.* **335**, 75–109 (1997).
- Zhou, Y., Wexler, A. & Wang, L.-P. Modelling turbulent collision of bidisperse inertial particles. *J. Fluid Mech.* **433**, 77–104 (2001).
- Reade, W. & Collins, L. Effect of preferential concentration on turbulent collision rates. *Phys. Fluids* **12**, 2530–2540 (2000).
- Wang, L. P. & Maxey, M. Settling velocity and concentration distribution of heavy particles in homogeneous isotropic turbulence. *J. Fluid Mech.* **256**, 27–68 (1993).
- Balkovsky, E., Falkovich, G. & Foufoula, A. Intermittent distribution of inertial particles in turbulent flows. *Phys. Rev. Lett.* **86**, 2790–2793 (2001).
- Brumfiel, G. How raindrops form. *Phys. Rev. Focus* **7**, story 14 (22 March 2001), <http://focus.aps.org/v7/st14.html>.
- Shraiman, B. & Siggia, E. Scalar turbulence. *Nature* **405**, 639–646 (2000).
- Falkovich, G., Gawedzki, K. & Vergassola, M. Particles and fields in fluid turbulence. *Rev. Mod. Phys.* **73**, 913–975 (2001).
- Pinsky, M., Khain, A. & Shapiro, M. Collision efficiency of drops in a wide range of Reynolds numbers. *J. Atmos. Sci.* **58**, 742–766 (2001).
- Johnson, D. B. The role of giant and ultragiant aerosol particles in warm rain initiation. *J. Atmos. Sci.* **39**, 448–460 (1982).
- Levin, Z., Wurzel, S. & Reisin, T. Modification of mineral dust particles by cloud processing and subsequent effects on drop size distribution. *J. Geophys. Res.* **105**, 4501–4512 (2000).
- Saffman, P. & Turner, J. On the collision of drops in turbulent clouds. *J. Fluid Mech.* **1**, 16–30 (1956).
- Raju, N. & Meiburg, N. The accumulation and dispersion of heavy particles in forced two-dimensional mixing layers. *Phys. Fluids* **7**, 1241–1264 (1995).
- Maxey, M. R. & Riley, J. J. Equation of motion for a small rigid sphere in a nonuniform flow. *Phys. Fluids* **26**, 883–889 (1983).
- Vekstein, G. *Physics of Continuous Media: A Collection of Problems with Solutions* 93–94 (Adam Hilger, Bristol, 1992).
- Grits, B., Pinsky, M. & Khain, A. Formation of small-scale droplet concentration inhomogeneity in a turbulent flow as seen from experiments with an isotropic turbulence model. *Proc. 13th Int. Conf. on Clouds and Precipitation* Vol. 1, 138–141 (Am. Met. Soc., Reno, 2000).
- Kostinski, A. & Shaw, R. Scale-dependent droplet clustering in turbulent clouds. *J. Fluid Mech.* **434**, 389–398 (2001).
- Brenguier, J.-L. & Chaumat, L. Droplet spectra broadening in cumulus clouds. *J. Atmos. Sci.* **58**, 628–641 (1999).

- Belin, F., Maurer, J., Tabeling, P. & Willaime, H. Velocity gradient distribution in fully developed turbulence: an experimental study. *Phys. Fluids* **9**, 3843–3850 (1997).
- Davila, J. & Hunt, J. C. R. Settling of small particles near vortices and in turbulence. *J. Fluid Mech.* **440**, 117–145 (2001).
- Crowe, C. T., Chung, J. N. & Troutt, T. R. *Particulate Two Phase Flow* Ch. 18 (ed. Roce, M. C.) 626 (Butterworth-Heinemann, Oxford, 1993).
- Marble, F. E. Dynamics of dusty gases. *Annu. Rev. Fluid Mech.* **2**, 397–461 (1970).

Acknowledgements

We thank A. Khain, V. Lebedev and M. Pinsky for discussions, and the Minerva and Israel Science Foundations for support.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.F.

(e-mail: finfal@wicc.weizmann.ac.il).

Macroecological patterns of phytoplankton in the northwestern North Atlantic Ocean

W. K. W. Li

Biological Oceanography Section, Bedford Institute of Oceanography, Department of Fisheries and Oceans, Dartmouth, Nova Scotia B2Y 4A2, Canada

Many issues in biological oceanography are regional or global in scope¹; however, there are not many data sets of extensive areal coverage for marine plankton. In microbial ecology, a fruitful approach to large-scale questions is comparative analysis^{2,3} wherein statistical data patterns are sought from different ecosystems, frequently assembled from unrelated studies⁴. A more recent approach termed macroecology characterizes phenomena emerging from large numbers of biological units by emphasizing the shapes and boundaries of statistical distributions, because these reflect the constraints on variation⁵. Here, I use a set of flow cytometric measurements to provide macroecological perspectives on North Atlantic phytoplankton communities. Distinct trends of abundance in picophytoplankton and both small and large nanophytoplankton underlaid two patterns. First, total abundance of the three groups was related to assemblage mean-cell size according to the 3/4 power law of allometric scaling in biology^{6,7}. Second, cytometric diversity⁸ (an taxonomic measure of assemblage entropy) was maximal at intermediate levels of water column stratification⁹. Here, intermediate disturbance shapes diversity through an equitable distribution of cells in size classes, from which arises a high overall biomass. By subsampling local fluctuations, macroecology reveals meaningful patterns of phytoplankton at large scales.

The data were collected from 23 oceanographic cruises spanning a 13-year period. We have published previously a description of the 14 earlier cruises (1989–98), a map of the stations, and sampling methods¹⁰. More recently (1999–2001), we re-occupied the stations on the Scotian shelf in spring and autumn, and the Labrador Sea in spring. Although stations were located across seven biogeochemical provinces¹, only measurements in the four northwestern ones, representing 82% of the data set, were included in the present analysis. These four contiguous provinces were Atlantic Arctic (ARCT), boreal polar (BPLR), northwest Atlantic shelves (NWCS) and Gulf Stream (GFST). The stations were in an area bounded approximately by 38° N, 61° N, 42° W and 67° W. The extracted data

set comprised 6,339 seawater samples collected from an average of 9 to 10 depths in the photic zone of 656 hydrocasts. Although large, this data set lacked any winter measurements.

These data sets provide a clear delineation of phytoplankton variability in the ocean afforded by thousands of measurements. In a plot of phytoplankton abundance against chlorophyll biomass, the boundaries and internal structure of the data domain are indicated by colour-coded density regions (Fig. 1a). A polygon to encompass the data is given by the unit contour line, which excludes apparent outliers. This polygon has two rising edges: the bottom edge was defined by spring phytoplankters, whereas the top edge was defined by autumn phytoplankters. Large cells were more prevalent in spring whereas small cells were more so in autumn. The polygon also has a falling top edge such that above approximately 0.5 mg chlorophyll per m^3 , maximum cell abundance decreased with increasing biomass, indicating the importance of large cells. This transition level of biomass may be the maximum potential for the picoplankton ($<2\text{ }\mu\text{m}$) component of the community¹¹. At higher levels of photoautotrophic biomass, a substantial contribution is expected from the microphytoplankton, which were not counted. However, the upper boundary of the polygon would be largely unaltered because microphytoplankton in these waters, mainly diatoms and dinoflagellates, occur at concentrations between 1 and 1,000 cells ml^{-1} , which are too low to make a significant difference.

The structure of flow cytometric signatures for individual samples can be expressed by an index of so-called cytometric

diversity⁸. In this approach, the classification of phytoplankton is based not on taxonomy, but on ataxonomic categories in the bivariate biooptical domain of cell size and chlorophyll content defined by flow cytometric light scatter and red fluorescence. As such, cytometric diversity indicates richness in physiological as well as genetic variations. The index of cytometric diversity is calculated as Hill's diversity number of order one¹²; in other words, it is the exponential Shannon–Wiener index of the cytometric categories ($\exp H'$). It is a measurement embodying the ideas of richness and evenness that define the uncertainty of an organism sampled at random¹³; it is also a measurement of widespread use in phytoplankton ecology⁹.

In the northwestern North Atlantic, the scatter plot of cytometric diversity against phytoplankton abundance filled a triangular polygon (Fig. 1b), indicating maximum diversity at an intermediate abundance of about 2,000 cells ml^{-1} . Phenomenologically, maximum diversity was a balance between richness and evenness. As more cells appeared in the community, diversity increased because new categories were added. However, as even more cells appeared, they were added to existing categories giving them dominant status, thus decreasing evenness and diversity.

A striking pattern was revealed in the manner in which different size classes were related to the total standing stock of chlorophyll *a* biomass (Fig. 2a). In the range of chlorophyll from about 10 to

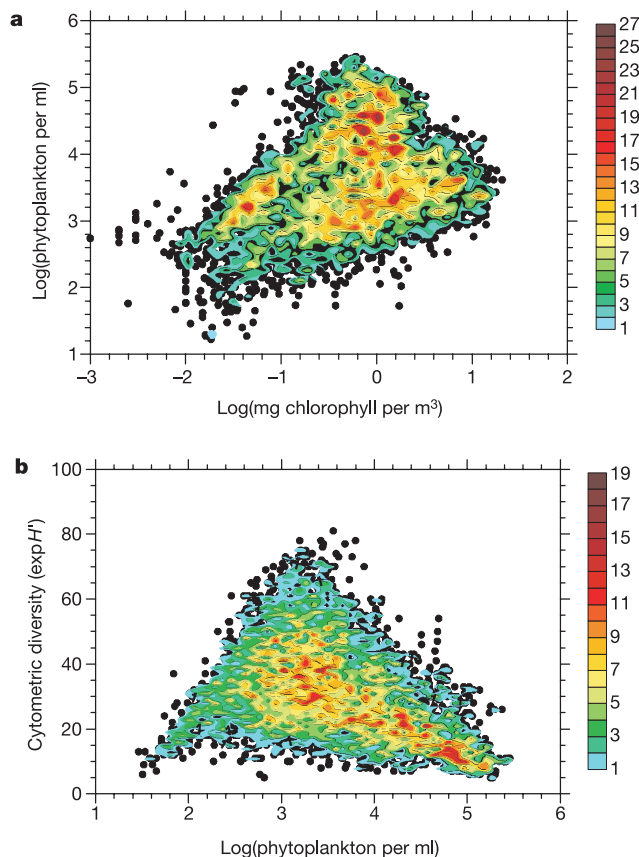


Figure 1 Range of phytoplankton variability in the northwestern North Atlantic Ocean. **a**, Phytoplankton abundance versus chlorophyll concentration on a volumetric basis. **b**, Cytometric diversity⁸ versus phytoplankton abundance. Samples were collected by Niskin bottles from depths of 0–100 m. The geographic distribution of the samples ($n = 6,339$) was 10% ARCT, 11% BPLR, 14% GFST and 65% NWCS. Contour levels indicate the number of overlapping data points in a grid region; the colour code indicates density. H' , Shannon–Wiener index.

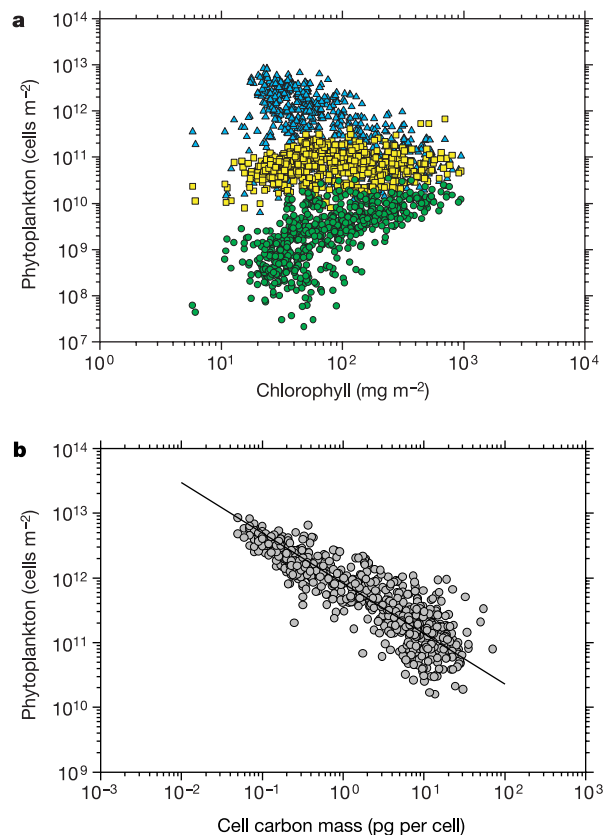


Figure 2 Relationship of phytoplankton abundance to biomass and mean cell mass.

a, Phytoplankton abundance versus chlorophyll concentration on an areal basis ($n = 656$). Picophytoplankton ($<2\text{ }\mu\text{m}$, blue triangles), small nanophytoplankton ($2\text{--}10\text{ }\mu\text{m}$, yellow squares) and large nanophytoplankton ($10\text{--}20\text{ }\mu\text{m}$, green circles) were classified by reference to forward light scatter. Areal values were calculated by trapezoidal integration of volumetric values from 0 to 100 m. **b**, Phytoplankton abundance on an areal basis versus cellular carbon mass. Cell carbon was calculated from cell volume¹⁹, which was based on the mean equivalent spherical diameter of the phytoplankton assemblage. The assemblage was an abundance-weighted average from 0 to 100 m depth. The line is the model II regression $\log N = 11.91 - 0.78 \log M_c$.

1,000 mg m⁻², a decrease in picoplankton (<2 µm) was mirrored by an increase in large nanoplankton (10–20 µm); however, small nanoplankton (2–10 µm) remained relatively invariant throughout. Here, the picoplankton were primarily *Synechococcus* cyanobacteria and small eukaryotic algae. There were no indications of *Prochlorococcus* cyanobacteria on the Scotian shelf or the Labrador Sea. The small nanoplankton were rich in 19'-hexanoyloxyfucoxanthin, indicating prymnesiophytes. The large nanoplankton were rich in fucoxanthin, indicating primarily diatoms but including other algae as well (Bouman *et al.*, manuscript in preparation). It is widely held that marine phytoplankton communities are assembled by adding larger cells to a relatively uniform background of smaller cells^{11,14,15}. The results here indicate that this uniform background is in fact the small nanoplankton. As phytoplankton communities grade from low to high biomass, there is a reduction in picoplankton, an increment in large nanoplankton, and no apparent net change in small nanoplankton.

These systematic patterns in the different size groups lead to phytoplankton communities that conform to an apparently universal relationship between population density and organism size¹⁶. The principles of allometric scaling^{6,7} proceed from a large body of observations that relate metabolic rate (*B*) to body mass (*M*) according to the 3/4 power law, $B = B_0 M^{3/4}$, where *B*₀ is a proportionality coefficient. From this, the allometric basis for population density of terrestrial plants can be interpreted as follows¹⁷. In individual plants, the rate of resource use (*Q*) is proportional to *B*, and thus $Q \propto M^{3/4}$. The number of individuals per unit area (*N*) that can be supported by a rate of resource supply (*R*) is $N = RQ^{-1}$.

It then follows that $N \propto M^{-3/4}$. Geometric mean model II regression¹⁸ of the North Atlantic phytoplankton data (Fig. 2b) indicates that $\log N = 11.91 - 0.78 \log M_C$, where *M*_C is the carbon mass per cell calculated from the mean equivalent spherical diameter of the phytoplankton assemblage using an allometric formula to convert cell volume to cell carbon¹⁹. The slope is indistinguishable from -0.75 ($r^2 = 0.77$, $n = 635$, 95% confidence interval = -0.74 to -0.81). This result greatly extends the generality of the 3/4 power law for phytoplankton beyond laboratory cultures of freshwater chlorophytes²⁰ and marine species in a coastal fjord of Sweden²¹.

In the ocean, phytoplankton community structure is shaped heavily by physical processes of advection and turbulence²². Recent work²³ points to the direct influence of mesoscale vertical motion on the slope of the phytoplankton size–abundance spectrum: upward motion retains large cells in the upper layer against their sinking tendency and therefore results in a flatter spectrum slope. In the limit of maximum upward water motion, the least negative slopes were about -0.75, in accord with the general law. By contrast, size–abundance spectra of entire plankton ecosystems that span across trophic levels have slopes close to -1 (ref. 24), indicating the loss of available energy with increasing body size in systems where small organisms are generally consumed by larger organisms²⁵.

The external (abiotic) control of assemblage structure was evident when population abundances of the three phytoplankton classes were examined in relation to the degree of mixing of the water column (Fig. 3a). Picoplankton were clearly favoured under conditions of high stratification, generally marked by low nutrients and high temperature. As stratification diminished, the balance shifted away from picoplankton and towards large nanoplankton. Under well-mixed conditions, no clear pattern emerged from the sparse data. Throughout the range of mixing conditions, the abundance of small nanoplankton was invariant. Together, these patterns seem to underlie the cytometric diversity (Fig. 3b). At an intermediate level of stratification, a clear maximum in diversity was evident where picoplankton abundance and large nanoplankton abundance converged, from opposing directions, to small nanoplankton abundance (Fig. 3a, b). If it can be agreed that mixing of the water column is a disturbance that has the potential to alter ecological equilibria, then the intermediate disturbance hypothesis^{26,27} becomes a useful basis to understand the role of allogenic forcing on processes of marine phytoplankton assembly. Essentially, strong disturbance excludes all but a few robust species whereas weak disturbance allows strong competition; diversity is therefore maximum when the level of disturbance is intermediate. Intermediate disturbance may be viewed as a factor shaping diversity (Fig. 3b) through an equitable distribution of cells in size classes (Fig. 3a), from which arises a high overall biomass (Fig. 2a). This is an idea that easily fits into the current view of how different regimes of turbulence and nutrients influence the prominent life forms of phytoplankton²². Somewhere in between the small-cell regime of low turbulence and low nutrients and the large-cell regime of high turbulence and high nutrients lies a regime of high diversity.

Reynolds's thesis that the principles discerned from pelagic microscopic vegetation, subject to scaling adjustments, have general applicability in ecology²⁷ is a suitable context in which to view the macroecological patterns of phytoplankton. The link now developing between concepts of microbial size spectra, rules of self-organization and hydrodynamic forcing^{23,28} is strengthened by the realization that picophytoplankton, small nanophytoplankton and large nanophytoplankton have distinct responses to abiotic control that can shape community structure. □

Methods

Chlorophyll *a* was measured in acetone extracts of particulate matter collected on Whatman GFF filters using a Turner Designs fluorometer. Phytoplankton in aliquots

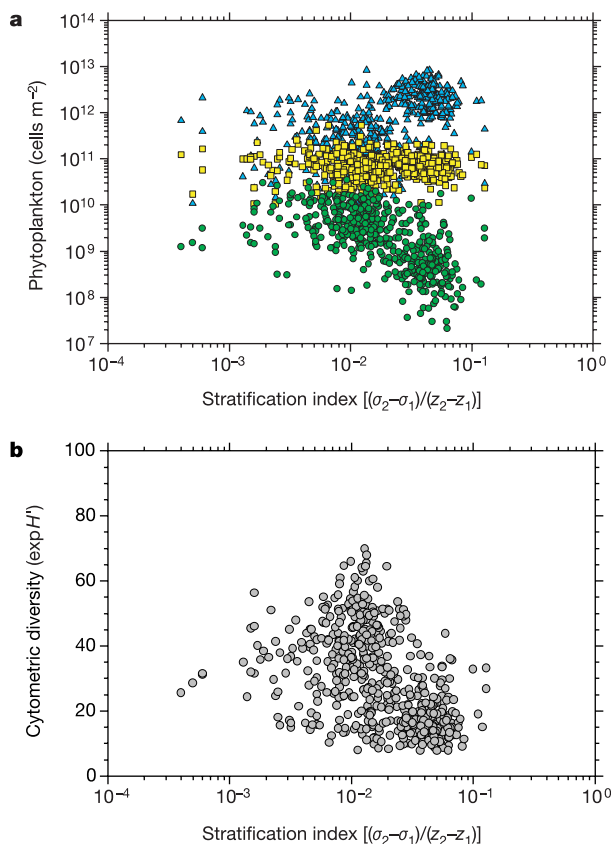


Figure 3 Influence of water-column stability on phytoplankton abundance (**a**) and cytometric diversity (**b**) ($n = 656$). The stratification index is the seawater density (σ) difference normalized to the depth (z) difference, which was about 90 m in most cases. Cytometric diversity was calculated as the abundance-weighted average from 0 to 100 m. Phytoplankton size classes are colour coded as in Fig. 2a.

(≤ 0.5 ml) of sea water were analysed by flow cytometry²⁹ using a Becton Dickinson FACS Sort. As population abundance is inversely related to cell size, detection of phytoplankton in the small aliquots was usually limited to cells that were abundant, namely the picophytoplankton ($< 2 \mu\text{m}$) and nanophytoplankton ($2\text{--}20 \mu\text{m}$). The microphytoplankton ($> 20 \mu\text{m}$) were not included in our measurements. This limitation is discussed elsewhere³⁰. Although the microphytoplankton contribute significantly to total photoautotrophic biomass at various times and places, they were seldom significant on a numerical basis. Cytometric forward light scatter was calibrated using synthetic microspheres²⁹ and therefore cannot be assumed to measure accurately the size of phytoplankton cells²⁸. However, the approximate positions of $2 \mu\text{m}$ and $10 \mu\text{m}$ on the light scatter scale were confirmed by flow cytometric analysis of the phytoplankton assemblage after physical size fractionation using Nuclepore polycarbonate membranes.

Received 23 May; accepted 11 July 2002; doi:10.1038/nature00994.

- Longhurst, A. *Ecological Geography of the Sea* (Academic, San Diego, 1998).
- Cole, J., Lovett, G. & Findlay, S. (eds) *Comparative Analyses of Ecosystems: Patterns, Mechanisms, and Theories* (Springer, New York, 1991).
- Gasol, J. M. & Duarte, C. M. Comparative analyses in aquatic microbial ecology: how far do they go? *FEMS Microb. Ecol.* **31**, 99–106 (2000).
- Li, W. K. W. Annual average abundance of heterotrophic bacteria and *Synechococcus* in surface ocean waters. *Limnol. Oceanogr.* **42**, 1746–1753 (1998).
- Brown, J. H. *Macroecology* (Univ. Chicago Press, Chicago, 1995).
- Peters, R. H. *The Ecological Implications of Body Size* (Cambridge Univ. Press, Cambridge, 1983).
- Brown, J. H. & West, G. B. (eds) *Scaling in Biology* (Oxford Univ. Press, New York, 2000).
- Li, W. K. W. Cytometric diversity in marine ultraphytoplankton. *Limnol. Oceanogr.* **42**, 874–880 (1997).
- Sommer, U., Padisák, J., Reynolds, C. S. & Juhász-Nagy, P. Hutchinson's heritage: the diversity-disturbance relationship in phytoplankton. *Hydrobiologia* **249**, 1–7 (1993).
- Li, W. K. W. & Harrison, W. G. Chlorophyll, bacteria and picophytoplankton in ecological provinces of the North Atlantic. *Deep-Sea Res. II* **48**, 2271–2293 (2001).
- Chisholm, S. W. in *Primary Productivity and Biogeochemical Cycles in the Sea* (eds Falkowski, P. G. & Woodhead, A. D.) 213–237 (Plenum, New York, 1992).
- Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
- Legendre, P. & Legendre, L. *Numerical Ecology*, 2nd edn (Elsevier, Amsterdam, 1998).
- Yentsch, C. S. & Phinney, D. A. A bridge between ocean optics and microbial ecology. *Limnol. Oceanogr.* **34**, 1694–1705 (1989).
- Ciotti, A. M., Lewis, M. R. & Cullen, J. J. Assessment of the relationships between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient. *Limnol. Oceanogr.* **47**, 404–417 (2002).
- Damuth, J. D. Common rules for animals and plants. *Nature* **395**, 115–116 (1998).
- Enquist, B. J., Brown, J. H. & West, G. B. Allometric scaling of plant energetics and population density. *Nature* **395**, 163–165 (1998).
- Laws, E. A. & Archie, J. W. Appropriate use of regression analysis in marine biology. *Mar. Biol.* **65**, 13–16 (1981).
- Montagnes, D. J. S., Berges, J. A., Harrison, P. J. & Taylor, F. J. R. Estimating carbon, nitrogen, protein, and chlorophyll *a* from volume in marine phytoplankton. *Limnol. Oceanogr.* **39**, 1044–1060 (1994).
- Agusti, S. & Kalf, J. The influence of growth conditions on the size dependence of maximal algal density and biomass. *Limnol. Oceanogr.* **34**, 1104–1108 (1989).
- Belgrano, A., Allen, A. P., Enquist, B. J. & Gillooly, J. F. Allometric scaling of maximum population density: a common rule for marine phytoplankton and terrestrial plants. *Ecol. Lett.* (in the press).
- Cullen, J. J., Franks, P. J. S., Karl, D. M. & Longhurst, A. in *Biological-Physical Interactions in the Sea* (eds Robinson, A. R., McCarthy, J. J. & Rothschild, B. J.) *The Sea Vol. 12* 297–336 (Wiley, New York, 2002).
- Rodríguez, J. et al. Mesoscale vertical motion and the size structure of phytoplankton in the ocean. *Nature* **410**, 360–363 (2001).
- Rodríguez, J. Some comments on the size-based structural analysis of the pelagic ecosystem. *Sci. Mar.* **58**, 1–10 (1994).
- Cyr, H. in *Scaling in Biology* (eds Brown, J. H. & West, G. B.) 267–295 (Oxford Univ. Press, New York, 2000).
- Connell, J. Diversity in tropical rain forests and coral reefs. *Science* **199**, 1304–1310 (1978).
- Reynolds, C. S. *Vegetation Processes in the Pelagic: a Model for Ecosystem Theory* (Ecology Institute, Oldendorf/Luhe, 1997).
- Cavender-Bares, K. K., Rinaldo, A. & Chisholm, S. W. Microbial size spectra from natural and nutrient enriched ecosystems. *Limnol. Oceanogr.* **46**, 778–789 (2001).
- Li, W. K. W. Composition of ultraphytoplankton in the central North Atlantic. *Mar. Ecol. Prog. Ser.* **122**, 1–8 (1995).
- Li, W. K. W. & Dickie, P. M. Monitoring phytoplankton, bacterioplankton, and viroplankton in a coastal inlet (Bedford Basin) by flow cytometry. *Cytometry* **44**, 236–246 (2001).

Acknowledgements

I thank all seagoing staff of the Biological Oceanography Section at Bedford Institute of Oceanography for collecting samples. I am grateful to C. Reynolds, G. Harrison, T. Platt and P. Falkowski for reading the manuscript. I also thank A. Belgrano for sharing his knowledge and pre-publication work on allometry with me. Financial support was provided by DFO Strategic Science Fund in the Ocean Climate Program, and the Interdepartmental Panel on Energy R&D.

Competing interests statement

The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to W.K.W.L. (e-mail: Lib@mar.dfo-mpo.gc.ca).

Coding of smooth eye movements in three-dimensional space by frontal cortex

Kikuro Fukushima*, Takanobu Yamanobe*, Yasuhiro Shinmei*, Junko Fukushima*, Sergei Kurkin* & Barry W. Peterson†

* Department of Physiology, Hokkaido University School of Medicine, West 7, North 15, Sapporo 060-8638, Japan

† Northwestern University Medical School, 303E Chicago Avenue, Chicago, Illinois 60611, USA

Through the development of a high-acuity fovea, primates with frontal eyes have acquired the ability to use binocular eye movements to track small objects moving in space¹. The smooth-pursuit system moves both eyes in the same direction to track movement in the frontal plane (frontal pursuit), whereas the vergence system moves left and right eyes in opposite directions to track targets moving towards or away from the observer (vergence tracking). In the cerebral cortex and brainstem, signals related to vergence eye movements—and the retinal disparity and blur signals that elicit them—are coded independently of signals related to frontal pursuit^{2–6}. Here we show that these types of signal are represented in a completely different way in the smooth-pursuit region of the frontal eye fields^{7–11}. Neurons of the frontal eye field modulate strongly during both frontal pursuit and vergence tracking, which results in three-dimensional cartesian representations of eye movements. We propose that the brain creates this distinctly different intermediate representation to allow these neurons to function as part of a system that enables primates to track and manipulate objects moving in three-dimensional space.

In two monkeys, we recorded the activity of 225 neurons that was modulated during frontal pursuit and/or vergence tracking of laser spots projected onto a vertical or horizontal screen (Methods). Of 122 neurons tested during both frontal pursuit and vergence tracking, 80 (66%) responded to both (three-dimensional (3D) tracking). Thirty (25%) responded only during frontal pursuit and 12 (9%) responded only during vergence tracking.

Of the 92 neurons that responded during vergence tracking, 39 were activated during divergence (Fig. 1a), 45 during convergence (Fig. 3a, f) and 8 during both (data not shown). The neuron shown in Fig. 1a–d was activated strongly as the monkey tracked a target moving away from him (Fig. 1a) and more weakly during downward pursuit (Fig. 1d). Because the horizontal screen used to present targets moving in depth was at nose level, divergence eye movements were accompanied by small (0.8°) upward eye movements. This combined motion that was required during vergence tracking could not explain the increased discharge of this neuron during divergence, because it was activated during downward pursuit.

Similar arguments could be applied to the 36 neurons that had no response to vertical pursuit, or a response in the wrong direction, to account for their observed modulation during vergence tracking. For neurons whose vertical pursuit sensitivity was in the correct direction to contribute to their modulation during vergence tracking, the contribution was too small to account for the modulation observed during vergence tracking. Figure 2a shows this for a divergence plus upward pursuit neuron whose eye velocity sensitivity at 0.3, 0.5 and 1.0 Hz was much smaller during frontal pursuit than during vergence tracking.

Figure 2b plots the relationship between vergence-tracking and frontal-pursuit (combined horizontal, vertical) sensitivities for five groups of neurons. Sensitivities varied widely with a tendency for

convergence neurons (open squares) to have higher sensitivities to both vergence tracking and frontal pursuit. The lack of any clear correlation between sensitivities to vergence tracking and frontal pursuit suggested that these two inputs to frontal eye fields (FEFs) were independent. Figure 2c shows the spatial relationship between frontal-pursuit and vergence-tracking sensitivities. The radius indicates the ratio of frontal-pursuit sensitivity to vergence-tracking sensitivity so that neurons primarily sensitive to vergence tracking are near the centre. The angle is the preferred direction of frontal pursuit. As expected, most neurons had a ratio of less than one, and preferred directions were distributed widely, as has been reported previously for frontal pursuit^{7–11}.

FEF neurons could show an apparent sensitivity to 3D tracking if they were sensitive to the motion of a single eye, as has been observed in brainstem neurons⁶. To rule out this possibility, we recorded responses during vergence tracking of targets moving in the midsagittal plane or planes aligned with the left and right eyes. None of the neurons that had vertical or oblique preferred directions showed a clear monocular preference (Fig. 1a–c)—modulation was similar regardless of whether the net vergence movement was produced by motion of the left, right or both eyes. Rate–velocity relationships for target movements in the three planes were also similar (Fig. 2a), again suggesting that none of these neurons had a monocular preference.

Summation also accounted for the behaviour of neurons that responded to both vergence tracking and horizontal frontal pursuit. Figure 3a and b shows the normalized discharge modulations (mean $\pm$ s.e.m.) of 10 neurons whose discharge increased during convergence and leftward pursuit (Fig. 3a, black and red lines, respectively). When the target moved in the left-eye-aligned plane, convergent movements of the right eye were accompanied by leftward frontal pursuit, and the discharge modulation during this condition (Fig. 3b, black) was predicted well by the linear addition of modulations related to vergence and horizontal tracking (Fig. 3b, blue) derived from data in Fig. 3a.

We made a similar comparison for 20 neurons that responded strongly to horizontal pursuit (the above 10 convergence plus leftward pursuit neurons, 5 convergence plus rightward pursuit neurons, 3 divergence plus leftward pursuit neurons, and 2 divergence plus rightward pursuit neurons). Figure 3c summarizes the comparison of actual and predicted modulations (gain and phase

relative to target velocity) of the 20 neurons during one-eye-aligned conditions in which modulation was in the additive direction. For most neurons, actual modulations during one-eye-aligned conditions were predicted well by addition of the two components, and the correlation coefficients were high ($r > 0.8$) with linear regression slopes close to one. Similar summation was also seen in the subtractive direction when off-saturation at zero firing rate was taken into account (data not shown).

Visual association cortex neurons may be sensitive to both the velocity of motion in the frontal plane and the disparity of the target, which signals its position in depth^{12–14}. During 3D sinusoidal target motion, such a neuron would discharge in phase with velocity of target motion in the frontal plane and with target position in depth. Most FEF neurons behaved differently. Modulation during 0.5-Hz frontal pursuit and vergence tracking was most closely related to velocity of target or eye motion (Figs 1a–c and 3a, d). Median phases (relative to velocity) of neurons that showed significant unidirectional responses were 10° lead and 1° lag during frontal pursuit and vergence tracking, respectively. Median sensitivities were 0.63 and 1.70 spikes s^{-1} per degree s^{-1} . Because modulations caused by frontal pursuit and vergence tracking add roughly linearly (Fig. 3a–c), we can show these sensitivities as 3D vectors in two views (Fig. 4a, b). Preferred directions are distributed widely in 3D space.

To check further that frontal-pursuit and vergence-tracking signals combine to produce enhanced tracking along oblique trajectories such as those shown in Fig. 4, we examined the activity of several neurons when the monkey tracked a virtual target (produced by dichoptic presentation of targets to right and left eyes in alternation using shutter glasses) that moved along a trajectory close to the predicted preferred oblique trajectory for that neuron. The normalized activity (mean $\pm$ s.e.m., Fig. 3e, black traces) of six neurons during pursuit along their expected optimal oblique directions was very close to the activity predicted (Fig. 3e, blue traces) by adding responses to the frontal-pursuit and vergence-tracking components of the trajectory (Fig. 3d). This indicates that it is indeed possible to predict a preferred direction of motion in 3D space from the component responses that we measured.

Recording locations in both monkeys were in the caudal part of the arcuate sulcus including the fundus and posterior bank, similar

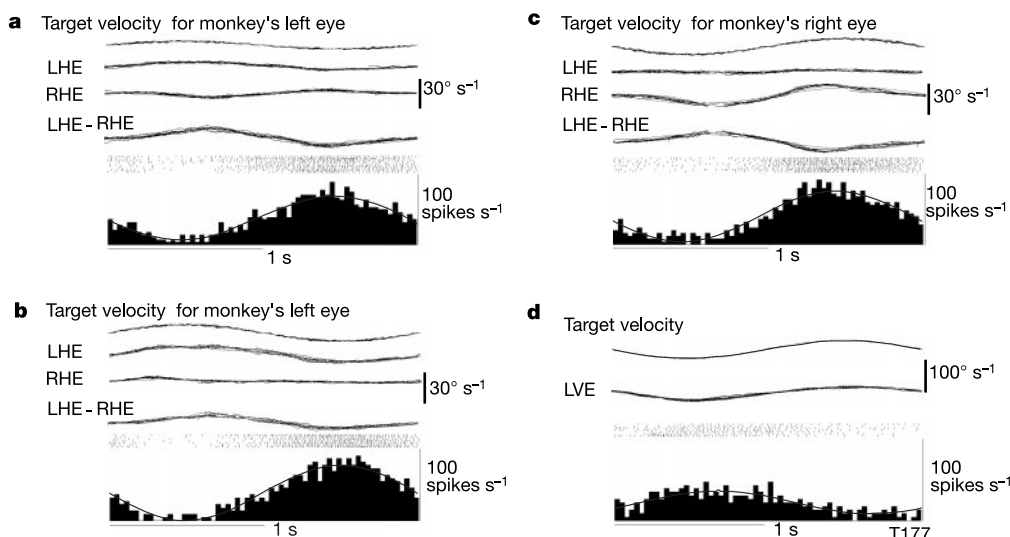


Figure 1 Representative activity of a pursuit neuron in the caudal frontal eye field. Response during target movement in the midsagittal (a), right-eye-aligned (b) and left-eye-aligned (c) planes and during vertical pursuit (d). LHE, left horizontal eye velocity;

RHE, right horizontal eye velocity; LVE, left vertical eye velocity. Vergence velocity is indicated as LHE – RHE. Velocity calibration is 30° s^{-1} (a–c) and 100° s^{-1} (d).

to those observed in previous studies (Fig. 4c and refs 7–11). Neurons responding to frontal pursuit, vergence tracking or both were intermixed. Injection of the GABA (γ -amino butyric acid) agonist muscimol ($10 \mu\text{g}$ in $1 \mu\text{l}$ of saline) at the site marked by an asterisk in Fig. 4c reduced the velocity of both frontal pursuit^{15,16}

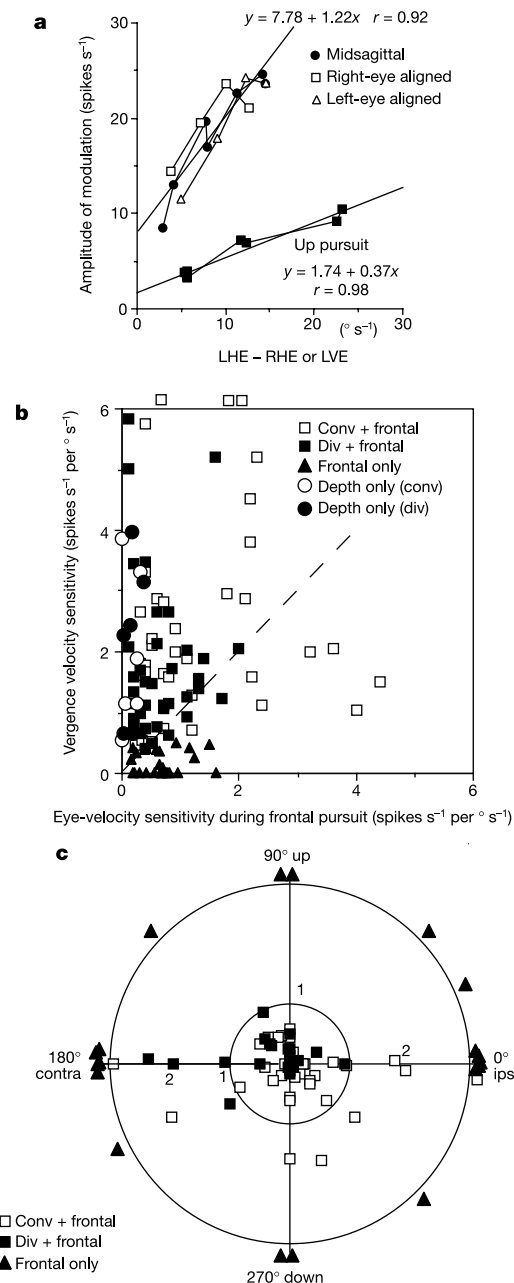


Figure 2 Eye velocity and vergence velocity sensitivity of caudal FEF neurons during frontal pursuit and vergence tracking. **a**, Discharge modulation of an up-pursuit plus divergence neuron plotted against vergence velocity for three target presentation conditions as indicated and vertical pursuit eye velocity with fitted linear regressions. LHE, left horizontal eye velocity; RHE, right horizontal eye velocity; LVE, left vertical eye velocity. **b**, Vergence-velocity sensitivity plotted against frontal-pursuit eye velocity sensitivity for the five indicated groups of neurons. The plot indicates some frontal-only and depth-only neurons with some eye and vergence velocity sensitivity, but harmonic distortion of their modulations were $>100\%$. **c**, Ratio of frontal pursuit to vergence tracking eye-velocity sensitivity plotted against preferred directions for frontal pursuit for the indicated three groups of neurons. Symbols outside the plot frame in **b** and **c** exceeded the plotting scale. In **c**, ipsi and contra indicate ipsilateral and contralateral to the recording side.

and vergence tracking by about 50%, which supports our conclusion that both movements are controlled by the same FEF region. This result also suggests that the correlation of FEF neural activity with 3D tracking reflects that FEF has a role in generating 3D smooth eye movements.

FEF signals related to frontal pursuit and vergence tracking are similar in other ways. When a target that the monkey was tracking was blanked for 400–800 ms (refs 10, 11) during frontal pursuit and vergence tracking, both eye movement and neural discharge continued in the absence of visual input (Fig. 3f). Thus, as for frontal pursuit^{10,11,17,18}, discharge related to vergence tracking is not

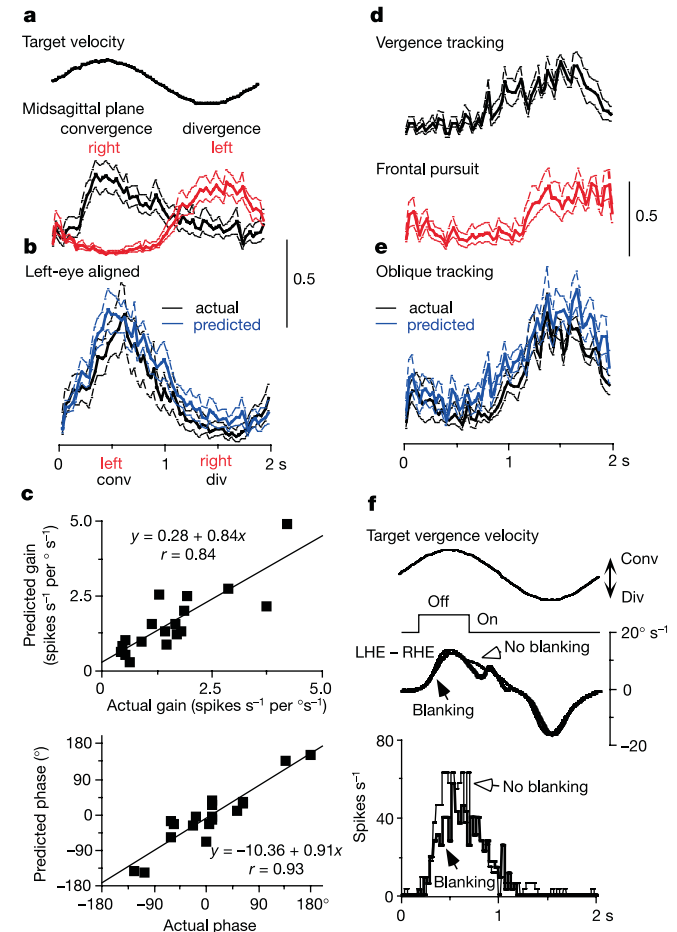


Figure 3 Linear summation of frontal pursuit and vergence tracking responses and maintenance of discharge during target blanking. **a**, **b**, Normalized discharge modulations (mean $\pm$ s.e.m.) during left-eye-aligned conditions for 10 convergent plus leftward pursuit neurons (**a**, actual) compared with predicted modulation (**b**, predicted) computed by adding modulations during midsagittal vergence tracking (**a**, black) and frontal pursuit (**a**, red). **c**, Comparison of actual and predicted modulations (gain and phase relative to target velocity) of 20 neurons with vergence plus horizontal pursuit sensitivities during one-eye-aligned conditions with fitted linear regressions. **d**, **e**, Normalized discharge modulations (mean $\pm$ s.e.m.) of six neurons during midsagittal vergence tracking (**d**, black), frontal pursuit (**d**, red) and when the monkey tracked a virtual target that moved along the expected optimal oblique directions for those neurons (**e**, black). Oblique target motion was presented by dichoptic spot stimulus for each eye. Predicted discharge modulation (**e**, predicted) was computed by adding modulations during vergence tracking and frontal pursuit (**d**). **f**, Eye movement and discharge of FEF neuron in vergence tracking cycles with (thick line) and without (thin line) target blanking. Note that tracking changes from the divergence direction to convergence in the absence of retinal input and that convergence-related activity of this neuron is identical on blanking and control trials until 300 ms into the blanking period. Afterwards activity during blanking is reduced and later still corrections occur after vision is restored.

dependent on retinal input. For vergence tracking, this also serves as a control that shows that responses do not simply reflect retinal disparity of a target.

The vergence and smooth-pursuit systems are considered to have separate neural substrates, as suggested more than a century ago^{1,19,20}. But 66% of our FEF neurons were sensitive to both frontal pursuit and vergence tracking. In the brainstem, vergence-related (not 3D) neurons located around the oculomotor nucleus project to the medial rectus subdivision and provide a vergence signal to these motoneurons^{1,4,5,21}. Neurons have also been reported whose discharge selectively reflects movement of one eye⁶, which may reflect precisely matched sensitivities to both frontal pursuit and vergence tracking. How could 3D motion signals in the FEF with their wide range of sensitivities be converted into commands to control the version and vergence motions of the two eyes? This could be accomplished by taking a weighted sum of the activity of neurons with different combinations of vergence and version activity. For example, consider a set of four neurons that respond with equal sensitivity to both frontal pursuit and vergence tracking, with D, C, L and R indicating, respectively, activation during tracking in divergence, convergence, leftward and rightward directions (so

that, for example, CL indicates a neuron that is activated equally during convergence and leftward tracking). The combination CL + CR – DL – DR will yield a net convergence (+)/divergence (–) signal. This could provide input to cortical^{2,5} or brainstem^{4,20} circuits that selectively control vergence and accommodation. CR + DR – CL – DL would provide a pure version signal. The combinations CR – DL and CL – DR yield monocular signals that are appropriate to drive the right and left eyes⁶. Selection of appropriate weights would allow such signals to be extracted from any set of neurons with a broad array of 3D preferred directions.

Because FEF discharge related to frontal pursuit and vergence tracking persists during target blanking, it is not dependent on sensory input but instead has a substantial motor or predictive component. Ultimately, however, the discharge and related eye movement are guided by visual input. It is interesting that 3D motion velocity signals have not yet been found in the visual system. Disparity-related signals reported in visual motion areas medial temporal (MT) and medial superior temporal (MST) cortices are mostly static depth position signals^{12,14} and discharge of MST neurons related to vergence velocity has been reported only in response to motion of large fields of dots²². Signals related to 3D

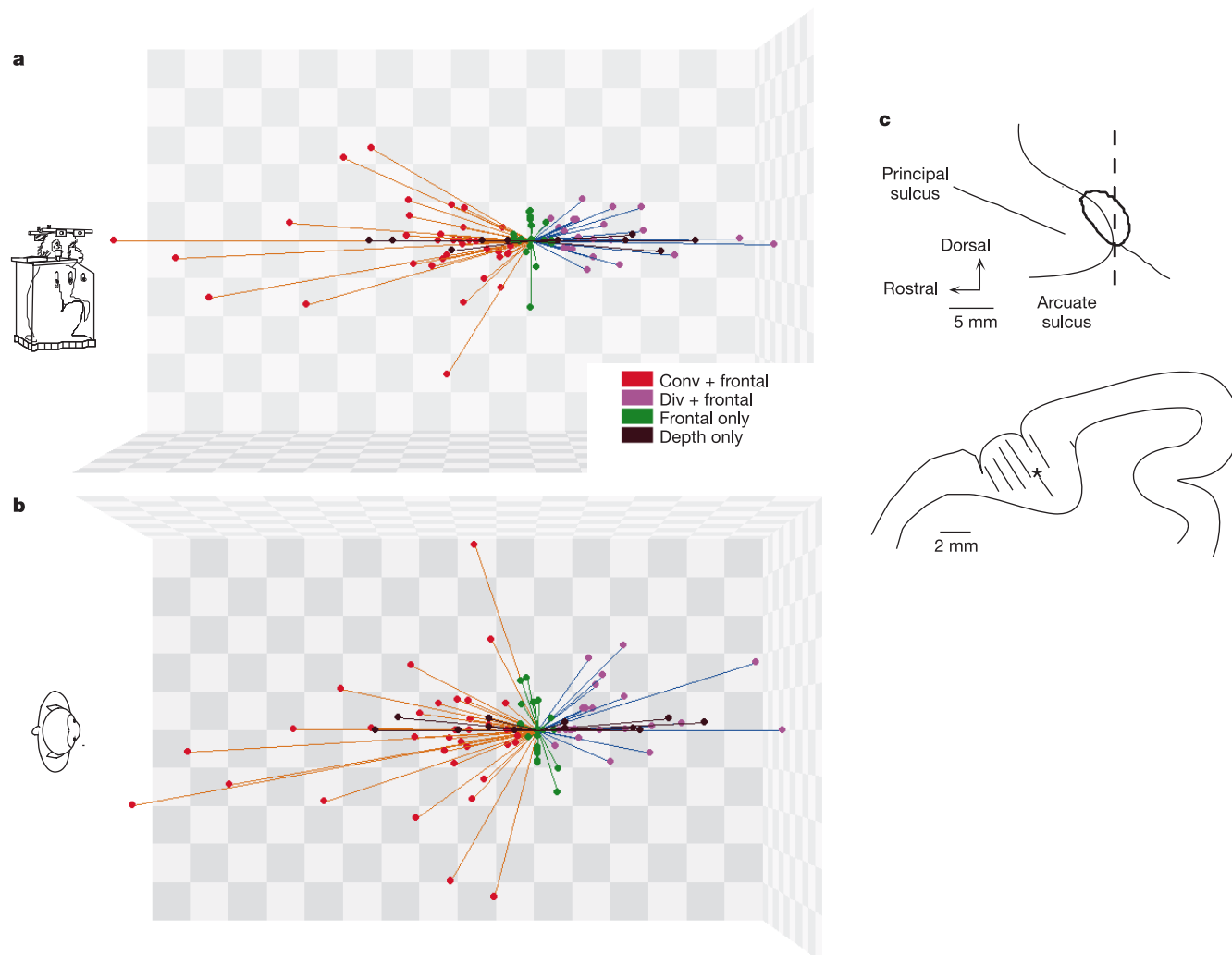


Figure 4 Three-dimensional sensitivity vectors of caudal FEF neurons and recording locations. **a, b**, 3D representation of preferred tracking directions and magnitudes of discharge sensitivity for the four indicated groups of FEF neurons (dots with lines). Three-dimensional sensitivity vectors were calculated for each neuron by assuming that frontal pursuit and vergence tracking sum linearly, and each direction was drawn to originate from a single point along the monkey's straight-ahead gaze. **c**, Recording locations in

monkey 1. The area indicated by the thick line in the top view of arcuate sulcus region indicates entry points of tracks in which neurons related to frontal pursuit and vergence tracking were observed. A cross-section through the area indicated by a dashed line shows trajectory of tracks containing such responsive neurons, which were typically encountered at depths of 3–5 mm from the surface.

target motion have been reported in the ventral intraparietal (VIP) area²³, but the activity depends on where the target trajectory will intersect the body, which suggests that position signals are also critical there. These findings underscore the uniqueness of the 3D motion velocity signals in FEF.

Thus, our data indicate that the brain creates an intermediate representation of tracking eye movements that are distinctly different from those reported in other sensory and motor areas of the brain. We propose that this occurred in response to an important factor driving primate evolution—the need to coordinate eye and hand movements to track and manipulate objects in a 3D world. By creating an intermediate representation of eye movements in a 3D coordinate frame that resembles the one in which we perceive the world, the brain opens the way for facile coordination of oculomotor signals with somatomotor signals coded in similar extrinsic frames. We predict that neighbouring frontal lobe areas will show related coding schemes for many different movements. Neurons in adjacent regions of FEF that control small saccades have disparity-tuned visual responses²⁴, which suggests that part of the saccadic FEF also uses a 3D coordinate frame. Conversely, largely separate coding of version and vergence movements in more rostral areas of FEF and area 8 has been reported (ref. 2). Perhaps these form part of a gaze-orienting system in which independent control of version and vergence is advantageous.

Recordings in the caudal ventral premotor cortex (area F4), which is located posterior to the smooth-pursuit area of FEF, have revealed neurons that code hand movements in a 3D extrinsic coordinate frame that is referred to the current centre of gaze²⁵. F4 neurons also exhibit sensory responses that reflect the 3D location of a visual target with respect to a point on the body surface^{26–29}. These signals could be readily derived by combining 3D eye and hand movement signals, whereas they would be more difficult to compute from signals intrinsic to each system such as version/vergence or joint-angle signals. □

Methods

Animals

Two monkeys (*Macaca fuscata*, 4.5 and 6.0 kg) were used in the experiments. We carried out all procedures in strict compliance with the Guide for the Care and Use of Laboratory Animals (DHEW Publication NIH85-23, 1985). Specific protocols were approved by the Animal Care and Use Committee of Hokkaido University School of Medicine. Methods for animal preparation, training and recording were similar to previous studies^{10,11} except for binocular recording and vergence task conditions, and are summarized here only briefly.

Animal preparation and training

A scleral search coil was implanted on each eye to record vertical and horizontal components of eye movement for both eyes. Monkeys' heads were restrained in the stereotaxic plane in a primate chair. Animals were trained to track a 0.2° laser spot backprojected onto a vertical screen 75 cm in front of the monkey's eyes for a reward of apple juice. A second horizontal screen was positioned at the level of the monkey's nose to project another 2 mm diameter laser spot from above. This screen was tilted 10° downward towards the front. From the monkey's viewpoint, the spot moved sinusoidally (0.3–1.0 Hz) from 28 to 70 cm in front of his eyes within the midsagittal, right- or left-eye-aligned plane. Accurate tracking of the midsagittal target required 1.3° movements of each eye, whereas tracking of right- or left-eye-aligned targets required 2.6° movement of the opposite eye. Because of the placement of the horizontal screen, vergence eye movements were accompanied by 0.8° vertical pursuit. A virtual target spot was presented on a computer monitor to test response of some neurons during 3D oblique tracking. For this, we used a time-multiplexed display with liquid crystal shutters synchronized to the alternating images for each eye at a refresh rate of 120 Hz.

Recordings

Extracellular recordings were made in the peri-arcuate sulcus region. Once an isolated neuron responding during frontal-pursuit or vergence tracking was encountered, the monkey was assigned vergence tracking tasks (Fig. 1a–c) followed by frontal-pursuit tasks using the vertical screen with a target moving at 0.5 Hz (5–10°) in different directions to determine the preferred direction for activation. We also tested frontal pursuit at smaller amplitudes (0.5–3.0°) for many neurons to confirm that neuronal discharge was related linearly to velocity of tracking.

Data collection and processing

As described previously¹¹, eye, target and chair position signals and their derivatives were

low-pass filtered (250 Hz) and digitized at 500 Hz. Neural discharge was discriminated and stored in temporal register with analog signals. Data segments containing saccades were identified by an interactive computer program and ignored in the following analysis. We constructed cycle histograms by averaging the discharge of each neuron over 10–30 cycles. Sinusoids were fit by least squares to neuron and eye movement responses. The signal-to-noise ratio of the response was defined as the ratio of amplitude of the fitted fundamental frequency component to the root mean-square amplitude of the third to the eighth harmonics. Harmonic distortion was defined as the ratio of the amplitude of the second harmonic to that of the fundamental. Responses with harmonic distortion >50% or signal to noise ratio <1.0 were discarded. Phase shifts were measured between the peak of the fundamental component of the response and the peak stimulus velocity. Phases of neural discharge relative to version or vergence eye-velocity were then calculated by subtraction.

We estimated preferred direction of individual neuron response during pursuit using a gaussian function as described^{10,11,30}. Each neuron's eye-velocity sensitivity during vergence tracking and frontal pursuit at 0.5 Hz was calculated by dividing the amplitude of the fundamental component of discharge modulation either by the amplitude of vergence tracking velocity or by the amplitude of eye-velocity modulation along the preferred frontal-pursuit direction. For neurons that had vertical pursuit sensitivity, we subtracted the modulation estimated to be due to the vertical component of eye movement from the discharge modulation during vergence tracking. (We chose sensitivity with respect to eye movement rather than sensitivity with respect to 'linear excursion of the target' in external space because we were relating FEF activity to eye movement. Choosing linear excursion would reduce the 2.7:1 sensitivity ratio in favour of vergence tracking that was described above by a factor of 6, resulting in a 0.44:1 sensitivity ratio in regard to target excursion.)

We normalized discharge modulations relative to maximal discharge rate during midsagittal vergence tracking and frontal pursuit for each neuron. Discharge modulations during one-eye-aligned conditions and 3D oblique tracking using a virtual display were also normalized relative to this maximal rate for each neuron. Normalized modulations were then averaged for many neurons to obtain population response.

To examine whether higher vergence-velocity sensitivity of many neurons was due to the effects of different target size (larger for convergence, smaller for divergence), we compared pursuit response using the vertical screen with a target of two different sizes (0.2° and 0.6° in diameter) comparable to the target sizes from the monkey's viewpoint at farthest and nearest distance on the horizontal screen. We also compared pursuit response to the target moving horizontally on the horizontal screen at the nearest and farthest positions. None of these conditions provided evidence that pursuit responses depended on target size or distance. Recording locations were histologically confirmed by making electrolytic lesions through the recording electrodes as described^{10,11}.

Received 26 April; accepted 13 June 2002; doi:10.1038/nature00953.

1. Leigh, R. & Zee, D. S. *The Neurology of Eye Movements* (Oxford Univ. Press, New York, 1999).
2. Gamlin, P. D. & Yoon, K. An area for vergence eye movement in primate frontal cortex. *Nature* **407**, 1003–1007 (2000).
3. Cumming, B. in *Visual Detection of Motion* (eds Smith, A. T. & Snowden, R. J.) 333–366 (Academic, London, 1994).
4. Mays, L. E., Porter, J. D., Gamlin, P. D. R. & Tellow, C. A. Neural control of vergence eye movements: neurons encoding vergence velocity. *J. Neurophysiol.* **56**, 1007–1021 (1986).
5. Mays, L. E. & Gamlin, P. D. Neuronal circuitry controlling the near response. *Curr. Opin. Neurobiol.* **5**, 763–768 (1995).
6. Zhou, W. & King, W. M. Premotor commands encode monocular eye movements. *Nature* **393**, 692–695 (1998).
7. MacAvoy, M. G., Gottlieb, J. P. & Bruce, C. J. Smooth pursuit eye movement representation in the primate frontal eye field. *Cereb. Cortex* **1**, 95–102 (1991).
8. Gottlieb, J. P., MacAvoy, M. G. & Bruce, C. J. Neural responses related to smooth pursuit eye movements and their correspondence with electrically elicited slow eye movements in the primate frontal eye field. *J. Neurophysiol.* **72**, 1634–1653 (1994).
9. Tian, J. & Lynch, J. C. Functionally defined smooth and saccadic eye movement subregions in the frontal eye field of Cebus monkeys. *J. Neurophysiol.* **76**, 2740–2771 (1996).
10. Tanaka, K. & Fukushima, K. Neuronal responses related to smooth pursuit eye movements in the periaruate cortical area of monkeys. *J. Neurophysiol.* **80**, 28–47 (1998).
11. Fukushima, K., Sato, T., Fukushima, J., Shinmei, Y. & Kaneko, C. R. S. Activity of smooth pursuit-related neurons in the monkey periaruate cortex during pursuit and passive whole body rotation. *J. Neurophysiol.* **83**, 563–587 (2000).
12. Maunsell, J. H. & van Essen, D. C. Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity. *J. Neurophysiol.* **49**, 1148–1167 (1983).
13. Roy, J. P., Komatsu, H. & Wurtz, R. H. Disparity sensitivity of neurons in monkey extrastriate area MST. *J. Neurosci.* **12**, 2478–2492 (1992).
14. Eifuku, S. & Wurtz, R. H. Response to motion in extrastriate area MST: disparity sensitivity. *J. Neurophysiol.* **82**, 2462–2475 (1999).
15. Shi, D., Friedman, H. R. & Bruce, C. J. Deficits in smooth pursuit eye movements after muscimol inactivation within the primate frontal eye field. *J. Neurophysiol.* **80**, 458–464 (1998).
16. Fukushima, K., Sato, T. & Fukushima, J. Vestibular-pursuit interactions: gaze-velocity and target-velocity signals in the monkey frontal eye fields. *Ann. NY Acad. Sci.* **871**, 248–259 (1999).
17. Ferrera, V. P. & Barborica, A. Predictive responses to invisible target motion in macaque frontal eye field. *Soc. Neurosci. Abstr.* **26**, 669 (2000).
18. Fukushima, K., Yamanobe, T., Shinmei, Y. & Fukushima, J. Predictive responses of peri-arcuate pursuit neurons to visual target motion. *Exp. Brain Res.* **145**, 104–120 (2002).
19. Semmlow, J. L., Weihong, Y. & Alvarez, T. L. Evidence for separate control of slow version and vergence eye movements: support for Hering's law. *Vision Res.* **38**, 1145–1152 (1998).
20. Gamlin, P. D. Subcortical neural circuits for ocular accommodation and vergence in primates. *Ophthalm. Physiol. Opt.* **2**, 81–89 (1999).
21. Judge, S. J. & Cumming, B. G. Neurons in the monkey midbrain with activity related to vergence eye movements and accommodation. *J. Neurophysiol.* **55**, 915–930 (1986).

22. Takemura, A., Inoue, Y., Kawano, K., Quaia, C. & Miles, F. A. Single-unit activity in cortical area MST associated with disparity-vergence eye movements: evidence for population coding. *J. Neurophysiol.* **85**, 2245–2266 (2001).
23. Colby, C. L., Duhamel, J. R. & Goldberg, M. E. Ventral intraparietal area of the macaque: anatomic location and visual response properties. *J. Neurophysiol.* **69**, 902–914 (1993).
24. Ferraina, S., Pare, M. & Wurtz, R. H. Disparity sensitivity of frontal eye field neurons. *J. Neurophysiol.* **83**, 625–629 (2000).
25. Kakei, S., Hoffman, D. S. & Strick, P. L. Direction of action is represented in the ventral premotor cortex. *Nature Neurosci.* **4**, 1020–1025 (2001).
26. Mushiaki, H., Tanatsugu, Y. & Tanji, J. Neuronal activity in the ventral part of premotor cortex during target-reach movement is modulated by direction of gaze. *J. Neurophysiol.* **78**, 567–571 (1997).
27. Fogassi, L. *et al.* Space coding by premotor cortex. *Exp. Brain Res.* **89**, 686–690 (1992).
28. Fogassi, L. *et al.* Coding of peripersonal space in inferior premotor cortex (area F4). *J. Neurophysiol.* **76**, 141–157 (1996).
29. Graziano, M. S., Hu, X. T. & Gross, C. G. Visuospatial properties of ventral premotor cortex. *J. Neurophysiol.* **77**, 2268–2292 (1997).
30. Krauzlis, R. J. & Lisberger, S. G. Directional organization of eye movement and visual signals in the floccular lobe of the monkey cerebellum. *Exp. Brain Res.* **109**, 289–302 (1996).

Acknowledgements

We thank B. Cumming, M. Goldberg, S. Lisberger, F. Miles and P. Strick for comments on the manuscript, and T. Akao and F. Sato for participation in some experiments. This work was supported in part by CREST of JST, Japanese Ministry of Education, Science, Culture and Sports, and Marna Cosmetics.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.E. (e-mail: kikuro@med.hokudai.ac.jp).

Loss of the *Lkb1* tumour suppressor provokes intestinal polyposis but resistance to transformation

Nabeel Bardeesy*, Manisha Sinha*, Aram F. Hezel*, Sabina Signoretti†, Nathaniel A. Hathaway*, Norman E. Sharpless*, Massimo Loda†, Daniel R. Carrasco† & Ronald A. DePinho*

* Department of Adult Oncology, Dana-Farber Cancer Institute and Departments of Medicine and Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA

† Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115 USA

Germline mutations in *LKB1* (also known as *STK11*) are associated with Peutz–Jeghers syndrome (PJS), a disorder with predisposition to gastrointestinal polyposis and cancer¹. PJS polyps are unusual neoplasms characterized by marked epithelial and stromal overgrowth but have limited malignant potential². Here we show that *Lkb1*^{+/-} mice develop intestinal polyps identical to those seen in individuals affected with PJS. Consistent with this *in vivo* tumour suppressor function, *Lkb1* deficiency prevents culture-induced senescence without loss of *Ink4a/Arf* or *p53*. Despite compromised mortality, *Lkb1*^{-/-} mouse embryonic fibroblasts show resistance to transformation by activated *Ha-Ras* either alone or with immortalizing oncogenes. This phenotype is in agreement with the paucity of mutations in *Ras* seen in PJS polyps^{3,4} and suggests that loss of *Lkb1* function as an early neoplastic event renders cells resistant to subsequent oncogene-induced transformation. In addition, the *Lkb1* transcriptome shows modulation of factors linked to angiogenesis, extracellular matrix remodelling, cell adhesion and inhibition of *Ras* transformation. Together, our data rationalize several features of PJS polyposis—notably its peculiar histopathological

presentation and limited malignant potential—and place *Lkb1* in a distinct class of tumour suppressors.

Compared with other hereditary tumour syndromes, PJS has several unusual features. Although germline mutations in *LKB1* are associated with a cancer-prone condition^{1,5}, the *LKB1* gene is very rarely mutated or epigenetically silenced in sporadic tumours^{6–8}. Heterozygosity for PJS is characterized by gastrointestinal polyps, but these polyps (hamartomas) possess low malignant potential and comprise disorganized non-dysplastic gastrointestinal mucosa with prominent branching smooth muscle components². In addition, although gastrointestinal carcinomas develop with increased frequency in individuals with PJS, it is not clear whether hamartomas are precursor lesions of these carcinomas².

To investigate these paradoxical features of PJS, we generated mice carrying a conditional *Lkb1* allele (Fig. 1a, b). Mice carrying either one copy of the null allele (*Lkb1*^{-/-}) or a functional, floxed allele deleted for the neomycin resistance (*neo*^r) gene (*Lkb1*^{lox}) were generated in crosses with *Elia-Cre*⁹ or *CAGG-Flpe*¹⁰ transgenic strains, respectively. Molecular analyses showed the expected recombinant alleles and absence of *Lkb1* protein in mouse embryonic fibroblasts (MEFs) after Cre-mediated excision (Fig. 1c–e and see below). *Lkb1*^{+/-} and *Lkb1*^{lox/+} offspring were born at expected mendelian frequencies and showed no gross abnormalities. Consistent with previous results¹¹, *Lkb1*^{-/-} mice had a lethal condition that manifested at about embryonic day (E) 8.5–11, with defects in vasculogenesis and placental development (data not shown).

We studied the tumour predisposition of these mice and found that 40 of 59 *Lkb1*^{+/-} and 0 of 65 wild-type mice presented with symptoms of gastrointestinal obstruction at an average age of 43 weeks (Fig. 2a). Autopsy of symptomatic *Lkb1*^{+/-} mice showed that polyps (between 1 and >15) were present throughout the gastrointestinal tract (Fig. 2b, c, and Supplementary Information Fig. 1), presenting as mucosal hamartomas with histological features mirroring those encountered in individuals with PJS or juvenile polyposis². Detailed histopathological analysis failed to discern dysplastic or adenomatous changes in 15 polyps examined, and none of the polyps showed mutations in *Ki-ras* (Methods). A single *Lkb1*^{+/-} mouse presented with a benign serous cystadenoma of the pancreas and two *Lkb1*^{+/-} mice showed asymptomatic, benign uterine epithelial tumours (data not shown), but otherwise an extensive histological survey of 20 *Lkb1*^{+/-} mice failed to identify additional neoplasms in other organs or abnormal mucocutaneous pigmentation up to an age of 45 weeks.

Hamartomas in individuals with PJS show loss of the wild-type *LKB1* allele in the epithelial compartment^{3,4,12}. Correspondingly, laser capture microdissection (LCM) and allele-specific polymerase

Table 1 Analysis of *Lkb1* in polyps

Tumour no.	LCM*	IHC*
1	Loss	Negative
2	Loss	Negative
3	Retention	Positive
4	Retention	Negative
5	Loss	ND
6	Retention	Negative
7	Retention	Negative
8	Retention	Positive
9	Retention	Positive
10	Retention	Positive
12	Retention	ND
13	Retention	Negative
14	ND	Negative
15	ND	Positive
16	ND	Negative
17	ND	Positive
Total	3/12	8/14

*Retention/loss denotes status of wild-type *Lkb1* allele. IHC, immunohistochemistry using an antibody against *Lkb1*; LCM, laser capture microdissection; ND, not determined or not informative.

chain reaction (PCR) analysis showed clear loss of the wild-type allele in the epithelial component of 3 of 12 polyps arising in *Lkb1*^{+/-} mice (Fig. 2d, polyps 1, 2 and 5); analysis of the stromal component of polyp 2 showed that wild-type *Lkb1* had been retained (Fig. 2d, lane S2). To extend this molecular analysis, we evaluated the cellular expression of Lkb1. Immunohistochemistry of normal intestinal epithelium using an antibody against Lkb1 identified a gradient of Lkb1 expression, which was highest in the crypt cells and lowest in the postmitotic differentiated compartment (Fig. 2e, region below broken line), and expression in the stromal cells of the submucosal layer (data not shown). Immunoreactivity against Lkb1 in sections of whole bowel or cultured MEFs was predominantly perinuclear and cytoplasmic (Fig. 2f, h, note absence of signal in the *Lkb1*^{-/-} MEFs). In 8 of 14 polyps, immunoreactivity against Lkb1 was absent in the polyp epithelium but consistently retained in the neighbouring stroma and normal crypt epithelial cells (Fig. 2e, f, and Table 1); notably, four of eight polyps that retained wild-type *Lkb1* alleles showed a loss of immunoreactivity that was consistent with epigenetic gene silencing. In the polyps that retained expression of Lkb1 (Fig. 2g), other means of inactivation such as point mutations or microdeletions might be operative, although tumour promotion by haploinsufficiency remains a formal possibility.

Given the robust expression of Lkb1 in MEF cultures (Fig. 1e), we used a somatic deletion approach in this model system to evaluate the biological role of Lkb1. Early passage (<passage 5) *Lkb1*^{-/-} and wild-type MEFs, cultivated under a 3T9 or 3T3 protocol, showed similar growth kinetics (Fig. 3a and data not shown). *Lkb1*^{-/-} MEFs possessed a small but consistent increase in G1 content ($6.1 \pm 1.9\%$) and reciprocal reduction in G2 ($7.4 \pm 1.8\%$; Supplementary Information Fig. 2a). At later passage (>9), *Lkb1*^{+/-} and wild-type cultures underwent growth arrest, whereas all six independently derived *Lkb1*^{-/-} cultures showed unabated growth after more than 40 population doublings (Fig. 3a). Reconstitution of wild-type Lkb1, but not a kinase-dead Lkb1 mutant, restored the

ability of early passage *Lkb1*^{-/-} cultures to undergo passage-induced growth arrest (Fig. 3b). Despite their immortal growth on serial passage, early and late passage *Lkb1*^{-/-} cells did not form colonies efficiently when seeded at very low densities (data not shown). This phenotype of density-dependent immortalization suggested that signals emanating from neighbouring cells might contribute to the immortalizing effects of Lkb1 deficiency.

The phenomenon of passage-induced senescence in MEFs has been attributed to the activation of a cellular stress response elicited by *in vitro* growth conditions^{13,14}, a process that depends on intact function of the p19^{Arf}/p53 and retinoblastoma (Rb) pathways¹⁵⁻¹⁸. Serially passaged *Lkb1*^{-/-} MEFs showed a prominent reduction in accumulation of p53, p19^{Arf} and p16^{Ink4a} as compared with wild-type control cultures cultivated in parallel (Fig. 3c, d), and a diminished expression of the p53-induced gene *p21*^{Cip1} (data not shown). Notably, even in late passage (>20) *Lkb1*^{-/-} MEFs, the p19^{Arf}/p53 pathway remained intact, as judged by p53 stabilization and growth arrest after exposure to various DNA-damaging agents (Fig. 3e and Supplementary Information Fig. 2c, d) or to ectopic expression of activated Ras, p19^{Arf} or c-Myc (Fig. 3f, g, and data not shown). Similarly, an intact Rb pathway in *Lkb1*^{-/-} cells was suggested by normal expression and phosphorylation of Rb (data not shown), normal S-phase re-entry kinetics on refeeding after serum starvation (Supplementary Information Fig. 2b), and intact G1 arrest after DNA damage or enforced expression of p19^{Arf} (Supplementary Information Fig. 2c-e). Together, these data suggest that loss of Lkb1 immortalizes MEFs by attenuating the culture-shock induction of the *Ink4a/Arf* locus without directly impairing the function of p53 or Rb. A relationship between *Lkb1* and *Ink4a/Arf* was further suggested by the inability of Lkb1 to induce growth arrest in *Lkb1*^{-/-} *Ink4a/Arf*^{-/-} cells (data not shown).

MEFs with an intact p19^{Arf}/p53 pathway are sensitive to Ras-induced premature senescence, whereas cells with compromised mortality pathways are readily transformed by activated Ha-ras (H-

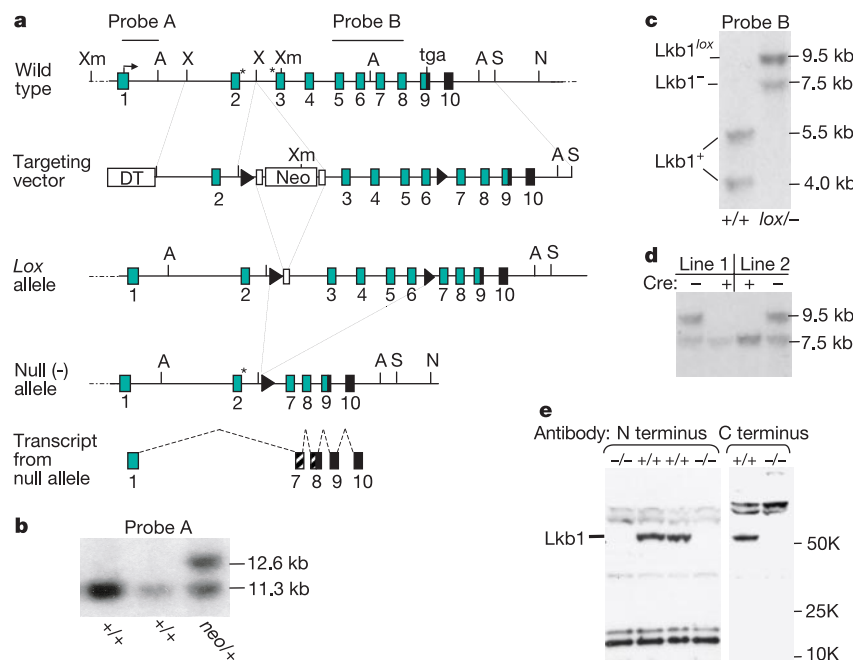


Figure 1 Targeting strategy and analysis of *Lkb1*. **a**, *Lkb1* genomic structure and recombinant alleles. The 3' untranslated region (black bars), and noncanonical splice sequences (asterisks) are indicated. The transcript from the *Lkb1* null allele eliminates exons 2-6, resulting in a translational frameshift (crosshatched bars). **b**, Southern analysis of DNA from ES cell clones showing targeting of the *Lkb1* locus. **c**, Southern

analysis of the *lox*, null (-), and wild-type (+) alleles. **d**, Southern analysis of DNA from *Lkb1*^{lox/-} MEFs, untreated (-) or infected with a Cre retrovirus (+). **e**, Western analysis of *Lkb1*^{lox/-} (-/-) and wild-type (+/+) MEFs after Cre expression, using antibodies to the amino (left) and carboxy (right) termini of Lkb1.

RASV12)^{15,19}. Unexpectedly, activated Ha-*ras* provoked premature senescence (that is, growth arrest and senescence-associated accumulation of β -galactosidase) in *Lkb1*^{-/-} cells, even at late passage (>20), despite their immortal phenotype and attenuated expression of p53 and p19^{Arf} (Fig. 3f and data not shown). In addition, the transformation efficiency of activated Ha-*ras* in combination with the SV40 large T antigen (T-Ag), dominant-negative p53 (*p53-DD*) or adenovirus E1a was severely reduced in *Lkb1*^{-/-} MEFs as compared with wild-type cultures (Fig. 4a, b). This resistance to transformation was also observed with T-Ag and a different dominantly acting oncogene, *Dbs*—a Dbl family guanine nucleotide exchange factor for Rho GTPases²⁰ (Fig. 4a, b). Furthermore, activated Ha-*ras* induced a marked decrease in growth of *Lkb1*^{-/-} *Ink4a/Arf*^{-/-} cells compared with wild-type *Ink4a/Arf*^{-/-} cells (Fig. 4c). This decrease in proliferation was restored by reintroduction of *Lkb1* with Ha-*ras* into *Lkb1*^{-/-} *ink4a/arf*^{-/-} cells (Fig. 4c). These data indicate that *Lkb1* deficiency unmasks a p19^{Arf}/p53-independent growth inhibitory pathway provoked by activated Ras.

The complex biological properties associated with *Lkb1*

deficiency prompted us to carry out transcriptional profiling of wild-type and *Lkb1*^{-/-} MEFs, as well as polyyps derived from *Lkb1*^{+/-} mice (Supplementary Information Figs 3 and 4). *Lkb1*^{-/-} MEFs and polyyps showed prominent alterations in the expression of several broad classes of genes, most notably secreted signalling molecules and regulators of the extracellular matrix (ECM). In particular, many genes that show substantially altered expression in *Lkb1*^{-/-} MEFs modulate the transformed phenotype of mesenchymal cells, which might explain the impaired transformation of *Lkb1*^{-/-} cells by activated Ha-*ras*. For example, lumican, fibulin-1D and Igfbp5 are established potent inhibitors of fibroblast transformation^{21–23}. *Lkb1*^{-/-} cells also show diminished expression of the *p8* gene, which encodes an HMG-I/Y-like (high mobility group) protein that is required for *ras* transformation of MEFs²⁴. Notably, increased expression of lumican and Igfbp5 in *Lkb1*^{-/-} cells remained despite enforced expression of activated *ras* (Fig. 5b). These alterations in *ras*-modulated genes suggest a molecular basis for the resistance of *Lkb1*^{-/-} MEFs to oncogene-induced transformation. Clearly, the absence of *Lkb1* does not result in a fundamental disruption of activated *ras* signalling because the

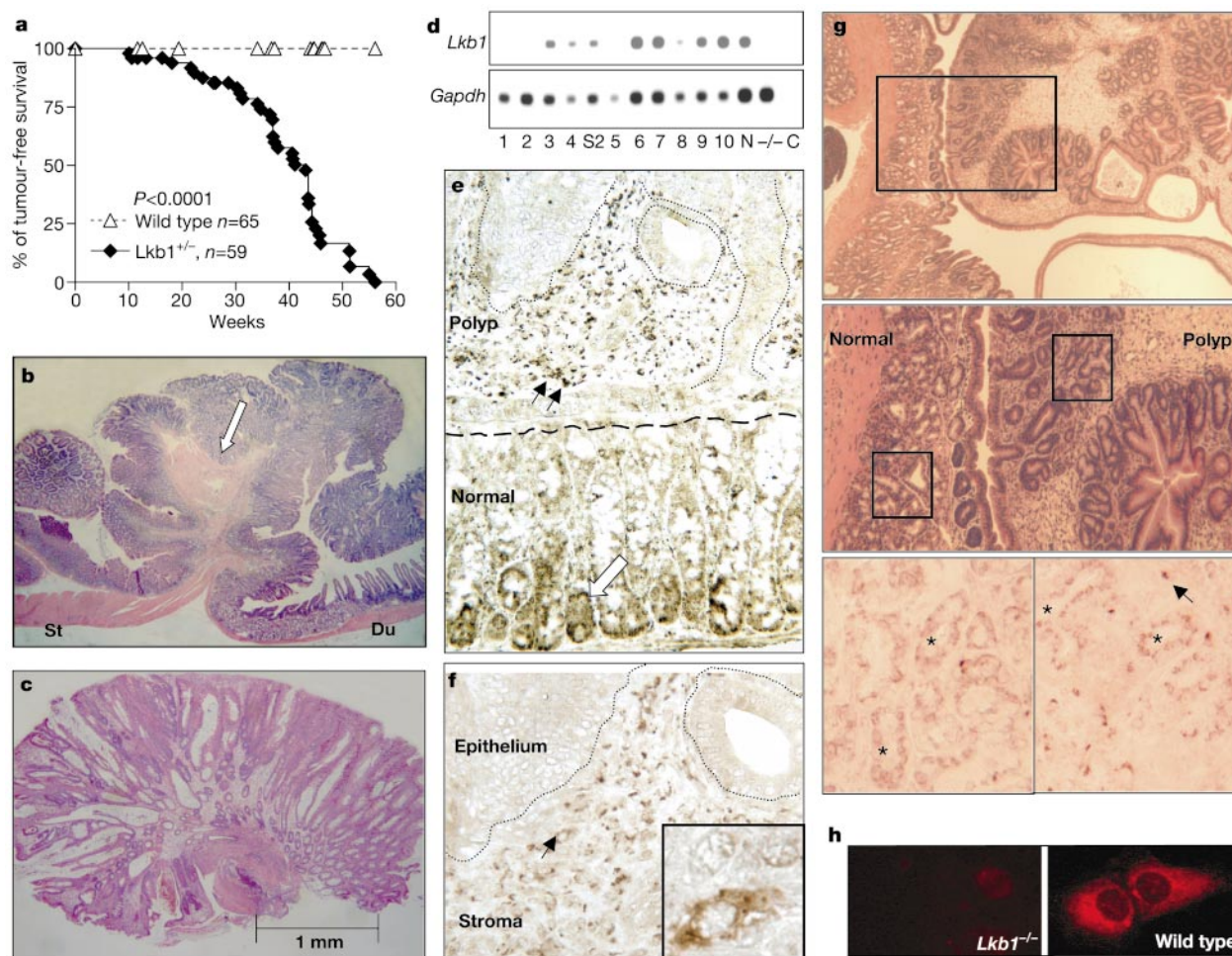


Figure 2 Gastrointestinal polyps in *Lkb1*^{+/-} mice. **a**, Tumour-free survival (Kaplan–Meier) analysis. **b**, **c**, Sections of a pyloric polyp with smooth muscle splaying (**b**, arrow) and a colonic polyp showing mucosal disorganization with cystically dilated glands (**c**) stained with haematoxylin and eosin. Du, duodenum; St, stomach. **d**, Allele-specific PCR analysis of DNA isolated by LCM from polyp epithelium (1–10) or stroma (S2) for the presence of wild-type *Lkb1* and *Gapdh* control alleles. N, *Lkb1*^{+/-} DNA; –/–, *Lkb1*^{-/-} DNA; C, no DNA. At longer exposure to film, the *Lkb1* allele was detectable at hypomolar ratios in polyps 1 and 2. **e**, **f**, Immunoreactivity against *Lkb1* is detected in normal colonic

epithelium (thick arrow) and polyp stroma (thin arrows) but is undetectable in the epithelium of polyp 2 (dotted lines). **f**, Higher magnification view of **e**. Inset, cytoplasmic and perinuclear staining of a stromal cell. **g**, Top, small intestine (left) and adjacent polyp 9 (right) stained with haematoxylin and eosin. The boxed region is shown at higher magnification in the middle. Immunoreactivity against *Lkb1* is detected (asterisks) both in the normal intestinal epithelium (bottom left) and the polyp epithelium (bottom right), as well as in the tumour stroma (arrow). **h**, Immunofluorescence of wild-type and *Lkb1*^{-/-} MEFs shows the specificity of the perinuclear and cytoplasmic staining pattern.

induction of a range of Ras targets such as Hb-Egf was comparable in wild-type and *Lkb1*^{-/-} MEFs (Fig. 5b and data not shown).

Expression profiling of a series of polyps compared with adjacent intestine also showed numerous reproducible changes in gene expression (Supplementary Information Fig. 4). Notably, several genes, including *Igfbp5*, *Pdgr-α* and *Mmp2*, showed marked elevation in both *Lkb1*^{-/-} MEFs and in polyps (Fig. 5c and not shown). Collectively, these datasets point to clear parallels between the MEFs and polyps to the extent that they highlight alterations in regulators of the cellular microenvironment and in pathways that influence *ras* transformation.

Given the preponderance of ECM regulators and secreted signaling molecules among the genes showing elevated expression in *Lkb1*^{-/-} cells, we wanted to determine whether the conditioned media from these cells could exert paracrine effects on gene expression. This issue is particularly relevant given the retention of *Lkb1* in the highly reactive stromal compartment of PJS polyps. Northern blot analysis showed that expression of MMP2 is increased in *Lkb1*^{-/-} MEFs regardless of the culture media

source (Fig. 5d). Conversely, cultivation of wild-type cells in the conditioned media from *Lkb1*^{-/-} cells produced a significant increase in MMP2 as compared with that detected after growth in conditioned media from wild-type MEFs; by contrast, *Igfbp5* was unaffected in this experiment (Fig. 5d).

In support of this observation, a marked increase in MMP2 could be seen in the stroma of polyps from *Lkb1*^{+/-} mice (data not shown). Consistent with a previous report¹¹, we also noted a fourfold increase in the amount of vascular endothelial growth factor (*Vegf*) in the conditioned media from *Lkb1*^{-/-} MEFs (data not shown), but no difference in steady-state levels of *Vegf* mRNA (Fig. 5a), suggesting an increased release of *Vegf* resulting from higher MMP activity in *Lkb1*^{-/-} cultures²⁵. These observations are in keeping with the idea that, *in vivo*, *Lkb1*-deficient cells may elicit alterations in signalling and ECM composition in the proximal stromal microenvironment.

Here we have modelled PJS intestinal polyposis and evaluated the biological and oncogenic impact of *Lkb1* deficiency. We show that *Lkb1*^{+/-} mice develop intestinal polyps that are indistinguishable from those in individuals affected with PJS including a complete absence of dysplastic changes and activating *ras* mutations. Paradoxically, although *Lkb1* deficiency confers immortal growth with retention of classical mortality pathways, *Lkb1*^{-/-} MEFs retain sensitivity to Ras-induced senescence and show resistance to the transforming effects of activated *ras* alone or in combination with an array of powerful immortalizing oncoproteins including T-Ag. This observation has added significance because, in contrast to

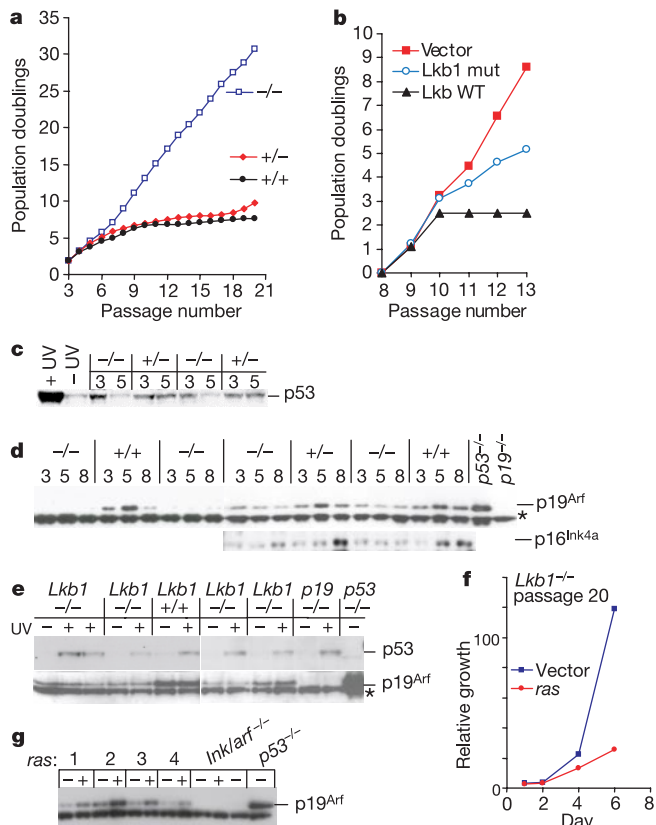


Figure 3 *Lkb1*^{-/-} MEF immortalization studies. **a**, Representative results of MEF cultures of the different *Lkb1* genotypes passaged on a 3T9 protocol. **b**, Growth of *Lkb1*^{-/-} MEFs after retroviral expression of wild type (WT) or the kinase-dead Lys78Ile *Lkb1* mutant (mut). **c**, Western blot analysis of p53 in serially passaged MEF lines. The *Lkb1* genotypes are indicated. Control lysates were irradiated (+UV) or untreated (-UV) MEF cultures. **d**, Western blot analysis of p19^{Arf} and p16^{Ink4a} in several serially passaged MEF lines. Asterisk denotes a nonspecific immunoreactive band. *p53*^{-/-} and *p16*^{Ink4a}/*p19*^{Arf}^{-/-} (labelled *p19*^{Arf}^{-/-}) MEFs are controls for immunoreactivity against p19^{Arf}. **e**, Intact *p19*^{Arf} and *p53* loci in late passage *Lkb1*^{-/-} MEFs. Top, induction of p53 by ultraviolet light (UV) in passage 20 *Lkb1*^{-/-} MEFs. The wild-type (+/+) lysate is from passage 14. Bottom, expression of p19^{Arf}. **f**, Activated Ha-ras causes growth arrest in passage 20 *Lkb1*^{-/-} MEFs. Data are representative of four MEF lines. **g**, Western blot showing induction of p19^{Arf} in passage 20 *Lkb1*^{-/-} MEFs lines infected with retroviruses encoding activated Ha-ras or empty vector (-).

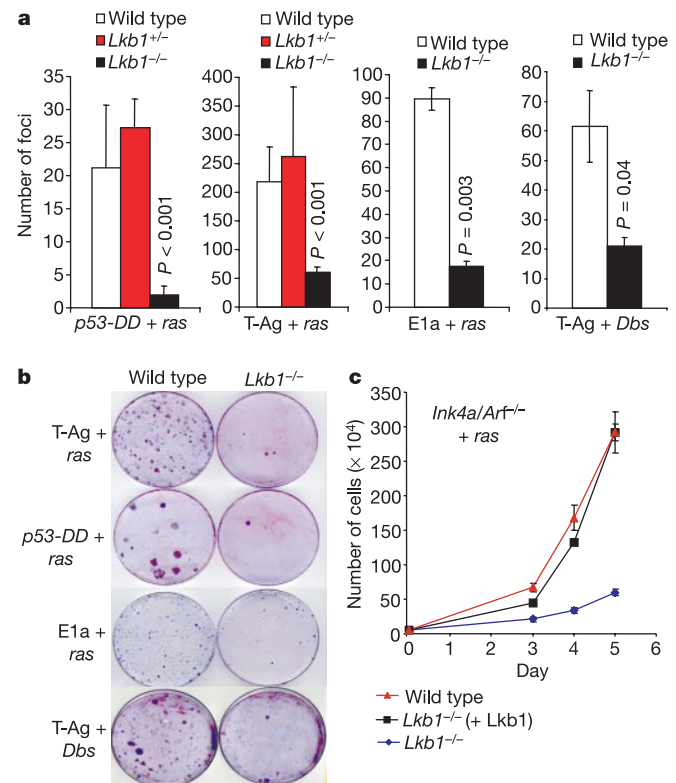


Figure 4 *Lkb1*^{-/-} cells are resistant to oncogenic transformation. **a**, Transformation assays. Values represent the mean number of transformed foci obtained for each *Lkb1* genotype in transformation assays with the indicated oncogenes. **b**, Representative plates stained with Giemsa from transformation assays. **c**, Growth curve of MEFs seeded at 50,000 per 10-cm dish after infection with Ha-ras. The genetic background of all cells was *Ink4a*/*Arf*^{-/-} and wild-type, *Lkb1*^{-/-} or *Lkb1*^{-/-} with reconstituted wild-type *Lkb1*.

other gastrointestinal neoplasms in humans, polyps arising in people affected with PJS rarely, if ever, carry oncogenic *ras* mutations^{3,4}.

In summary, *Lkb1* has a complex role in growth and transformation. Outwardly, the resistance to activated *ras* transformation and impaired malignant progression may seem difficult to reconcile with the polyposis and carcinoma-prone condition provoked by *Lkb1* deficiency. On further consideration, this paradoxical picture is consistent with the notable clinical features of PJS and its distinctive cancer genetics. Although individuals with PJS have a significantly increased risk of developing carcinomas, the PJS polyps themselves rarely show features of dysplasia². Our data indicate that there may be limitations on the type of initiating and/or cooperating events that are required for the malignant transformation of *LKB1*^{-/-} cells, despite the immortalizing effects of loss of *LKB1*. The extremely rare inactivation of *LKB1* in tumours other than those of individuals carrying germline mutations may also point to a constrained genetic context wherein *LKB1* loss enables malignant transformation⁶⁻⁸. Given the increased incidence of carcinoma associated with germline *LKB1* mutations, an additional implication of our studies is that the order of mutational events may be an important parameter in dictating the type of cooperating mutations and the malignant potential of the initiated neoplasm: early loss of *LKB1* may promote strictly benign neoplasia, whereas loss of *LKB1* in a later stage lesion could facilitate malignant progression. We therefore propose that *LKB1* is a context-dependent tumour suppressor gene, whose loss of function facilitates evasion of a classical barrier to neoplasia (senescence) and engenders a tumour-like stromal environment, but renders cells resistant to transformation by powerful oncogenic combinations.

Note added in proof: After this manuscript was submitted, Myoshi

*et al.*²⁹ and Jishage *et al.*³⁰ reported gastrointestinal polyps in *Lkb1* heterozygous mice; in contrast to our detection of *Lkb1* loss in a subset of polyps, these studies reported retention of *Lkb1* in the polyps analysed (2 and 3 polyps, respectively). □

Methods

Targeting construct, colony generation and genotyping

We cloned and mapped the *Lkb1* locus from a bacterial artificial chromosome library. The targeting vector (Fig. 1a) carried a negative selection marker for diphtheria toxin (DT), a positive selection marker for neomycin acetyltransferase (Neo), Frt sites (white rectangles) and *loxP* sites (black triangles). The restriction sites were *XmnI* (Xm), *XbaI* (X), *AvrII* (A), *SmaI* (S), *NotI* (N). We electroporated TC1 embryonic stem (ES) cells and selected transformed cells by standard techniques. We screened 92 clones by Southern analysis using a *XmnI* restriction enzyme and a 5' fragment external to the targeting construct (Fig. 1a, b, probe A) to identify nine recombinants. Blastocyst injections were carried out with three different targeted clones, and transmitting chimaeric mice were bred from CAGG-*Flpe* and *Elaa-Cre* transgenic mice^{9,10} to generate the *Lkb1*^{lox/-} and *Lkb1*^{lox/+} alleles, respectively (Fig. 1a). *Cre*⁺ *Lkb1*^{+/+} and *Flpe*⁺ *Lkb1*^{lox/+} males were backcrossed one or two times to FVB/n females and progeny of these matings that were *Cre*⁻ or *Flpe*⁻ were then backcrossed to littermates to yield the experimental cohort. Mice were genotyped by Southern analysis and multiplex PCR (primers and conditions are available from R.A.D. on request). Colonies were observed three times per week for morbidity, and autopsied for overt tumour or illness. Stool specimens were monitored weekly for occult blood using Hemoccult slides (Beckman Coulter).

Cellular analysis

We generated MEFs from 13.5 post-coitum embryos and grew them in DMEM medium plus 10% fetal calf serum (Hyclone), 50 μ M β -mercaptoethanol, penicillin and streptomycin. As *Lkb1*^{-/-} embryos did not develop past 11.5 days post coitum, *Lkb1*^{-/-} MEFs were produced by *in vitro* excision of the *Lkb1*^{lox} allele using a self-excising Cre expressing retrovirus²⁶. *Lkb1*^{lox/-} and littermate control *Lkb1*^{lox/+} and wild-type cultures were exposed to Cre retrovirus at passage 2 and found to sustain 95–100% deletion of the *Lkb1*^{lox} allele in short-term and 100% deletion in long-term passage cultures as determined by western blot, Southern blot and PCR analysis (Fig. 1d and data not shown). The structure of the transcript arising from the *Lkb1* null allele (Fig. 1a) was determined by PCR with reverse transcription and sequence analysis. For 3T9 or 3T3 analysis, 3×10^5 cells were passaged into six-well or 6-cm dishes, every 3 d. At least five independent lines were assayed per genotype, with three independent cultures per line. Senescence was determined as described²⁷. Growth curves were generated by seeding 25,000 cells per well in 12-well plates, each line in triplicate. We fixed plates on the indicated day, stained them with crystal violet, extracted them with 10% acetic acid, and measured the relative cell number at an absorbance of 595 nm. For S-phase re-entry experiments, passage 5 MEFs at 80% confluency were serum starved for 72 h and then stimulated by subculturing into serum-containing medium. Retroviral transduction, low-density seeding, SA- β -galactosidase, Ras-induced growth arrest and transformation assays were done as described²⁷. Transient transfection efficiencies did not differ between wild-type and *Lkb1*^{-/-} cells, as determined using green fluorescent protein. We used retroviruses for the *Ela-ras* transformation assay and for expression of *Lkb1*.

Molecular analysis

Cell lysates from MEFs were prepared and resolved on polyacrylamide gels as described²⁷. Western blots of *Lkb1* were carried out using antibodies D19 (Santa Cruz) or 07-093 (Upstate), or using 1G, a rabbit polyclonal antibody raised against the carboxy terminus of *Lkb1*; other proteins were blotted as described²⁷. Loading was assessed using α -tubulin (Sigma). For p53 induction, lysates were collected 12 h after exposure to ultraviolet radiation (100 μ J M⁻²). Immunohistochemistry was carried out on paraffin-embedded sections as described²⁸ except that antigen retrieval was carried out in Tris-EDTA, pH 6.8, at 90 °C for 30 min using antibody D19. We measured the immunofluorescence of *Lkb1* as described²⁸, using antibody D19 or 1G.

Expression profiling was carried out using the Affymetrix U74-A chip, and data were analysed using Affymetrix GeneChip 3.1 software. We isolated RNA from two pairs of littermate *Lkb1*^{-/-} and wild-type MEF cultures at passage 5. Expression changes were considered significant only if the GeneChip software made the call of 'different' and if the expression differences were greater than twofold in both sets of MEFs. Profiling of polyps versus adjacent normal intestine was done on four paired sets of specimens. Expression analysis was also carried out using nylon cDNA arrays according to the manufacturer's instructions (arrays MM001, MM006, MM009 and MM010; Superarray).

For LCM, paraffin-embedded 5- μ m sections were removed from paraffin by xylene, washed with ethanol, rehydrated in deionized water and stained with haematoxylin and eosin. Epithelial and stromal cells from the polyps and normal epithelial cells from the adjacent mucosa were microdissected using the PixCell II Laser Capture Microdissection System (Acturus Engineering). About 500–1,000 cells were digested in 30 μ l of buffer containing 10 mM Tris, 1 mM EDTA, 1% Tween 20 with 1 mg ml⁻¹ proteinase K at 37 °C for 12 h. Samples were heated to 94 °C for 10 min and then centrifuged. We used 3 μ l of the supernatant used as template for PCR (20 cycles) with primers specific for the *Lkb1* wild-type allele or to *Gapdh* and subjected the products to Southern blotting. LCM was repeated for all specimens to verify the results. For Ki-ras mutational analysis, exons 1 and 2 of Ki-ras were amplified by PCR from LCM material and sequenced.

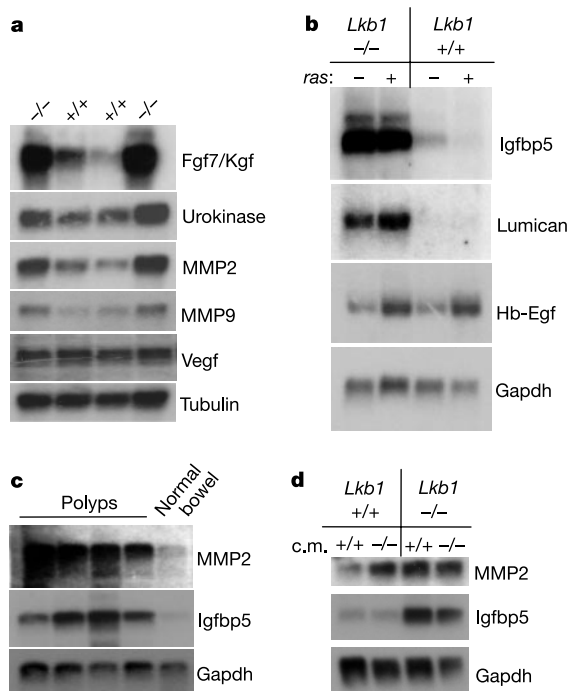


Figure 5 *Lkb1*^{-/-} expression profile. **a**, Northern blot analyses of candidate regulators of angiogenesis and extracellular matrix remodelling in wild-type and *Lkb1*^{-/-} MEFs. **b**, Expression of modulators of Ha-ras transformation in wild-type and *Lkb1*^{-/-} MEFs. Cells were either transduced with retroviruses expressing Ha-ras (H-RASV12, +) or with empty vector (-). **c**, Northern blot analysis of gastrointestinal polyps and normal bowel from *Lkb1*^{+/+} mice. **d**, Conditioned media from *Lkb1*^{-/-} cells induces MMP2 expression in *Lkb1*^{+/+} MEFs. Wild-type and *Lkb1*^{-/-} MEFs were cultured in conditioned media (c.m.) from either genotype and analysed by northern blot.

Received 5 April; accepted 10 July 2002; doi:10.1038/nature01045.

- Hemminki, A. *et al.* A serine/threonine kinase gene defective in Peutz–Jeghers syndrome. *Nature* **391**, 184–187 (1998).
- Cooper, H. S. *Pathology of the Gastrointestinal Tract* (eds Ming, S.-C. & Goldman, H.) 819–853 (Williams & Wilkins, Baltimore, 1998).
- Entius, M. M. *et al.* Molecular genetic alterations in hamartomatous polyps and carcinomas of patients with Peutz–Jeghers syndrome. *J. Clin. Pathol.* **54**, 126–131 (2001).
- Gruber, S. B. *et al.* Pathogenesis of adenocarcinoma in Peutz–Jeghers syndrome. *Cancer Res.* **58**, 5267–5270 (1998).
- Giardiello, F. M. *et al.* Very high risk of cancer in familial Peutz–Jeghers syndrome. *Gastroenterology* **119**, 1447–1453 (2000).
- Avizienyte, E. *et al.* Somatic mutations in LKB1 are rare in sporadic colorectal and testicular tumors. *Cancer Res.* **58**, 2087–2090 (1998).
- Avizienyte, E. *et al.* LKB1 somatic mutations in sporadic tumors. *Am. J. Pathol.* **154**, 677–681 (1999).
- Esteller, M. *et al.* Epigenetic inactivation of LKB1 in primary tumors associated with the Peutz–Jeghers syndrome. *Oncogene* **19**, 164–168 (2000).
- Lakso, M. *et al.* Efficient *in vivo* manipulation of mouse genomic sequences at the zygote stage. *Proc. Natl Acad. Sci. USA* **93**, 5860–5865 (1996).
- Rodriguez, C. I. *et al.* High-efficiency deleter mice show that *FLPe* is an alternative to *Cre-loxP*. *Nature Genet.* **25**, 139–140 (2000).
- Ylikorkala, A. *et al.* Vascular abnormalities and deregulation of VEGF in Lkb1-deficient mice. *Science* **293**, 1323–1326 (2001).
- Hemminki, A. *et al.* Localization of a susceptibility locus for Peutz–Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nature Genet.* **15**, 87–90 (1997).
- Ramirez, R. D. *et al.* Putative telomere-independent mechanisms of replicative aging reflect inadequate growth conditions. *Genes Dev.* **15**, 398–403 (2001).
- Sherr, C. J. & DePinho, R. A. Cellular senescence: mitotic clock or culture shock? *Cell* **102**, 407–410 (2000).
- Kamijo, T. *et al.* Tumor suppression at the mouse *INK4a* locus mediated by the alternative reading frame product p19^{ARF}. *Cell* **91**, 649–659 (1997).
- Serrano, M. *et al.* Role of the *INK4a* locus in tumour suppression and cell mortality. *Cell* **85**, 27–37 (1996).
- Sage, J. *et al.* Targeted disruption of the three *Rb*-related genes leads to loss of G₁ control and immortalization. *Genes Dev.* **14**, 3037–3050 (2000).
- Dannenberger, J. H., van Rossum, A., Schuijff, L. & te Riele, H. Ablation of the retinoblastoma gene family deregulates G₁ control causing immortalization and increased cell turnover under growth-restricting conditions. *Genes Dev.* **14**, 3051–3064 (2000).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and p16^{INK4a}. *Cell* **88**, 593–602 (1997).
- Whitehead, I. P. *et al.* Dependence of DbpA and Dbs transformation on MEK and NF-κB activation. *Mol. Cell. Biol.* **19**, 7759–7770 (1999).
- Yoshioka, N. *et al.* Isolation of transformation suppressor genes by cDNA subtraction: lumican suppresses transformation induced by v-src and v-K-ras. *J. Virol.* **74**, 1008–1013 (2000).
- Qing, J. *et al.* Suppression of anchorage-independent growth and matrigel invasion and delayed tumour formation by elevated expression of fibulin-1D in human fibrosarcoma-derived cell lines. *Oncogene* **15**, 2159–2168 (1997).
- Reeve, J. G., Guadano, A., Xiong, J., Morgan, J. & Bleehen, N. M. Diminished expression of insulin-like growth factor (IGF) binding protein-5 and activation of IGF-I-mediated autocrine growth in simian virus 40-transformed human fibroblasts. *J. Biol. Chem.* **270**, 135–142 (1995).
- Vasseur, S. *et al.* p8 is critical for tumour development induced by rasV12 mutated protein and E1A oncogene. *EMBO Rep.* **3**, 165–170 (2002).
- Bergers, G. *et al.* Matrix metalloproteinase-9 triggers the angiogenic switch during carcinogenesis. *Nature Cell Biol.* **2**, 737–744 (2000).
- Silver, D. P. & Livingston, D. M. Self-excising retroviral vectors encoding the Cre recombinase overcome Cre-mediated cellular toxicity. *Mol. Cell* **8**, 233–243 (2001).
- Sharpless, N. E. *et al.* Loss of p16^{INK4a} with retention of p19^{Arf} predisposes mice to tumorigenesis. *Nature* **413**, 86–91 (2001).
- Carrasco, D., Weih, F. & Bravo, R. Developmental expression of the mouse *c-rel* proto-oncogene in hematopoietic organs. *Development* **120**, 2991–3004 (1994).
- Miyoshi, H. *et al.* Gastrointestinal hamartomatous polyposis in Lkb1 heterozygous knockout mice. *Cancer Res.* **62**, 2261–2266 (2002).
- Jishage, K. *et al.* Role of Lkb1, the causative gene of Peutz–Jegher's syndrome, in embryogenesis and polyposis. *Proc. Natl Acad. Sci. USA* **99**, 8903–8908 (2002).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank L. Ritchie and J. Horner of DFCI mouse core for advice and assistance; S. Dymecki, H. Westphal, M. Oren, S. Lowe, D. Silver, C. Der & J. DeCaprio for advice and reagents; and D. Livingston, J. DeCaprio and W. Kaelin for comments on the manuscript. N.B. is supported by the ACS John Peter Hoffman Award and the Liss Family Fund grant for research in pancreatic cancer. R.A.D. is an American Cancer Society Professor and recipient of the Steven and Michele Kirsch Foundation Investigator Award. This work was supported by grants from the NCI (National Cancer Institute) and ACS (American Cancer Society).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.D. (e-mail: ron_depinho@dfci.harvard.edu).

SINAT5 promotes ubiquitin-related degradation of NAC1 to attenuate auxin signals

Qi Xie*, Hui-Shan Guo*, Geza Dallman*, Shengyun Fang†, Allan M. Weissman† & Nam-Hai Chua‡

* Laboratory of Molecular Cell Biology, Temasek Life Sciences Laboratory, National University of Singapore, 1 Research Link, 117604 Singapore

† Regulation of Protein Function Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA

‡ Laboratory of Plant Molecular Biology, Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

The plant hormone indole-3 acetic acid (IAA or auxin) controls many aspects of plant development, including the production of lateral roots^{1–3}. Ubiquitin-mediated proteolysis has a central role in this process. The genes *AXR1* and *TIR1* aid the assembly of an active SCF (Skp1/Cullin/F-box) complex that probably promotes degradation of the AUX/IAA transcriptional repressors in response to auxin^{4–8}. The transcription activator NAC1, a member of the NAM/CUC family of transcription factors, functions downstream of TIR1 to transduce the auxin signal for lateral root development⁹. Here we show that SINAT5, an *Arabidopsis* homologue of the RING-finger *Drosophila* protein SINA, has ubiquitin protein ligase activity and can ubiquitinate NAC1. This activity is abolished by mutations in the RING motif of SINAT5. Over-expressing SINAT5 produces fewer lateral roots, whereas over-expression of a dominant-negative Cys49 → Ser mutant of SINAT5 develops more lateral roots. These lateral root phenotypes correlate with the expression of NAC1 observed *in vivo*. Low expression of NAC1 in roots can be increased by treatment with a proteasome inhibitor, which indicates that SINAT5 targets NAC1 for ubiquitin-mediated proteolysis to downregulate auxin signals in plant cells.

To investigate NAC1 action, we carried out yeast two-hybrid assays using NAC1 as a bait. One protein that interacted with NAC1 showed extensive sequence homology to two C3HC4 RING-finger proteins, SINA from *Drosophila*¹⁰ and SIAH from human¹¹. Because the gene encoding this protein is located on chromosome 5 we designated this protein SINA of *Arabidopsis thaliana* 5 (SINAT5, AF480944; Fig. 1a). The identities between SINAT5 and SINA (M38384) and SIAH (U76247) are 33% and 36%, respectively.

When SINAT5 was used as the bait, it interacted with the *Arabidopsis* AtUBC9A (AF480945; a protein that shows homology to members of the yeast Ubc4/5 and human UbcH5 families), ubiquitin (data not shown) and with itself (Fig. 1b). The carboxy-terminal region was responsible for SINAT5 dimerization (data not shown), which is consistent with results reported for SIAH^{12,13}. We verified the dimerization of SINAT5 and interaction of the dimer with NAC1 by *in vitro* pull-down assays (Fig. 1c, d).

We determined the expression profile of SINAT5 by generating transgenic plants carrying a fusion of the SINAT5 promoter and the β-glucuronidase gene (*GUS*). SINAT5–*GUS* was expressed in low amounts in the vascular tissues of mature roots (Fig. 1e). But on treatment with auxin, expression was also detected in lateral root initials and in the elongation zone of the main root (Fig. 1e). This root expression pattern is very similar to that of NAC1 after induction by auxin (ref. 9 and Fig. 1e), which suggests that SINAT5 and NAC1 function in the same types of cell. As for NAC1, expression of SINAT5 was induced by auxin but more slowly (Fig. 1f).

To investigate the subcellular localization of SINAT5, we constructed a fusion of the green fluorescent protein (GFP) gene and

SINAT5 placed under the control of a CaMV 35S promoter. Transient expression in onion epidermal cells showed that, whereas GFP was distributed in both the cytosol and the nucleus, GFP-SINAT5 was localized predominantly in the nucleus (Fig. 1g), as has been reported for NAC1 (ref. 9).

As several proteins containing RING motifs function as E3 ubiquitin ligases *in vitro*^{14–16}, we determined whether SINAT5 also has E3 activity. We expressed SINAT5 in *Escherichia coli* as a fusion with maltose-binding protein (MBP). In the presence of ubiquitin, E1 and E2, purified MBP-SINAT5 could carry out self-ubiquitination, whereas control purified MBP was inactive (Fig. 2a). As negative controls, we constructed two mutant proteins, Cys49 → Ser (C49S) and His64 → Tyr (H64Y), in which the RING domain

was disrupted (Fig. 1a). Neither mutant protein could undergo self-ubiquitination (Fig. 2b), suggesting that an intact RING motif is needed for E3 activity. In addition, when present in a 5–10-fold excess the C49S mutant protein inhibited the self-ubiquitination activity of wild-type SINAT5 (Fig. 2c). A 10-fold increase in amounts of E2 did not reduce this inhibition (data not shown), which indicated that the E2 was not being sequestered by the excess amount of the C49S mutant present. These results show that the SINAT5 E3 activity requires dimerization and that the inactive C49S SINAT5 mutant can act as a dominant-negative protein, probably by forming inactive dimers with wild-type SINAT5.

The interaction between SINAT5, an E3 ligase, and NAC1 suggests that the SINAT5 may ubiquitinate the NAC1 *in vivo* and target it for proteasomal degradation. To show directly that NAC1 is a substrate of SINAT5, we carried out ubiquitination experiments using Myc-NAC1, which was immunoprecipitated from transgenic plants overexpressing the tagged protein. Myc-NAC1 was ubiquitinated by SINAT5 and this reaction was clearly dependent on E1, E2 and SINAT5 acting as an E3 ligase (Fig. 2d).

In transgenic plants, the overexpression of NAC1 produces more lateral roots, whereas the underexpression of NAC1 produces fewer lateral roots⁹. If SINAT5 ubiquitinates NAC1 *in vivo*, thereby promoting its degradation, we would expect overexpression of wild-type SINAT5 to reduce and overexpression of the dominant-negative C49S mutant to increase the numbers of lateral roots, respectively. These results were observed in transgenic experiments (Fig. 3). In addition, the main root length of SINAT5-overexpressing plants was longer than that of wild-type plants, whereas the opposite was true for C49S mutant plants. This effect of SINAT5 is similar to that observed in the auxin-response mutant *axr-1* (ref. 1). In a separate experiment, 10-day-old seedlings ($n = 30$) were transferred from MS (Murashige and Shooog) medium to fresh MS medium containing 2 μ M 1-naphthaleneacetic acid (1-NAA) for

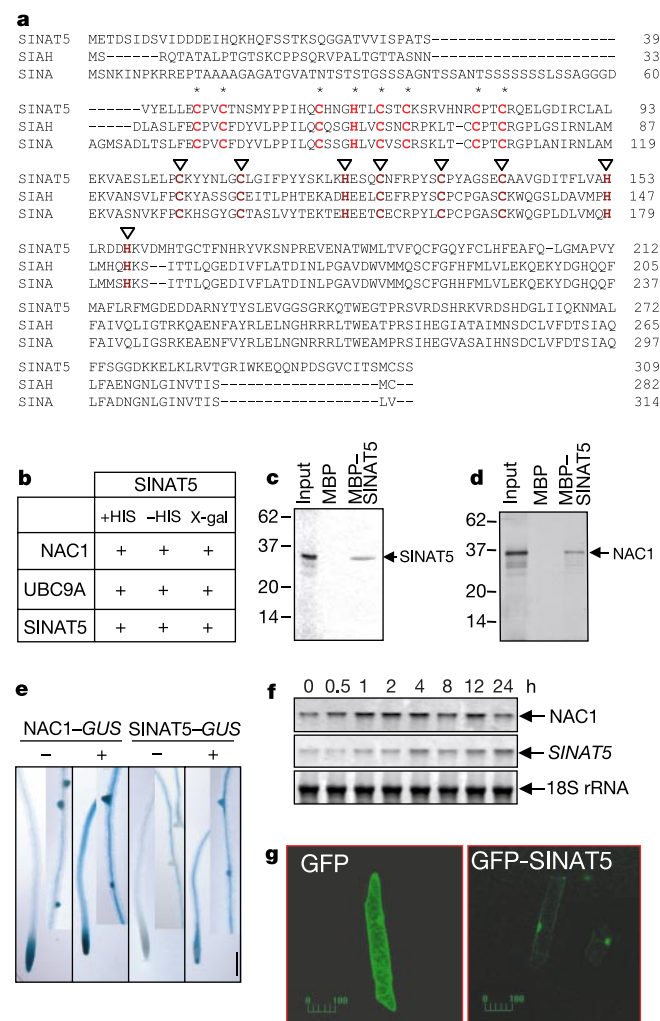


Figure 1 Molecular and cellular properties of SINAT5. **a**, Amino acid sequence comparison of SINAT5, SINA and SIAH. Asterisks and triangles indicate conserved cysteine and histidine residues, respectively, in the RING motif and the two zinc-fingers¹³. **b**, SINAT5 interacts with NAC1, AtUBC9A and itself. '+' indicates cell growth and β -galactosidase activity. **c**, Dimerization of SINAT5 *in vitro*. Recombinant MBP or MBP-SINAT5 was mixed with *in vitro* translated SINAT5 (Methods). **d**, SINAT5 binds to NAC1 *in vitro*. Recombinant MBP or MBP-SINAT5 was mixed with *in vitro* translated NAC1. **e**, Expression pattern of *SINAT5* and *NAC1*. Two-week-old transgenic seedlings were treated with (+) or without (–) 2 μ M 1-NAA for 6 h before being processed for GUS staining. Scale bar, 0.1 mm. **f**, Auxin induces expression of both *NAC1* and *SINAT5*. Cultured roots from wild-type seedlings⁹ were treated with 2 μ M 1-NAA, and the RNA expression at different times was analysed. **g**, SINAT5 is localized predominately in nuclei. Cells were analysed by confocal microscopy after 16 h of incubation. Scale bar, 1 μ M.

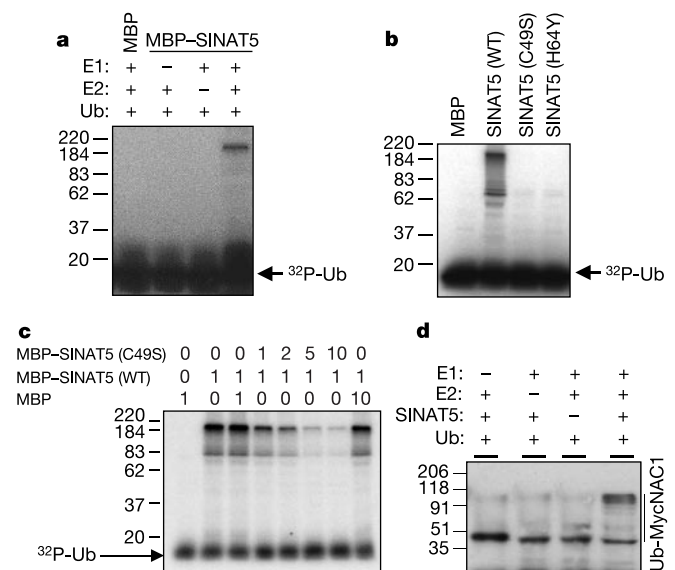


Figure 2 E3 activity of wild-type and mutant SINAT5. **a**, MBP-SINAT5 was assayed for E3 activity in the presence of E1, E2 and ³²P-labelled ubiquitin (Ub) as indicated. **b**, MBP-SINAT5 E3 activity is dependent on its RING motif. MBP-SINAT5 and mutants were assayed for self-ubiquitination. **c**, SINAT5 (C49S) blocked the E3 activity of SINAT5. Numbers indicate the relative amount of proteins present in the ubiquitination reaction, where 1 equals 100 ng of MBP or MBP-SINAT. **d**, Ubiquitination of NAC1 by SINAT5. Myc-NAC1 immunoprecipitated from transgenic plants was assayed for *in vitro* ubiquitination. Samples were processed for western blot analysis using an antibody against Myc.

48 h. The mean $\pm$ s.d. number of lateral roots per centimetre of main root was 13.4 ± 0.4 for the vector control, 7.2 ± 0.5 for plants overexpressing SINAT5, and 19.7 ± 0.5 for plants overexpressing the C49S mutant. These results show that SINAT5 inhibits auxin-induced lateral root development, whereas the C49S mutant promotes this process.

The expression of NAC1 in roots is low but can be increased considerably by the proteasome inhibitor MG 132 (Fig. 4a), which suggests that the low amount of protein detected is probably caused by continuous ubiquitin-mediated proteolysis. Pull-down assays using extracts prepared from transgenic plants showed that both wild-type SINAT5 and the C49S mutant could interact with NAC1 in plant cells (Fig. 4b). Consistent with the post-translational regulation of NAC1 by SINAT5 *in vivo*, overexpression of wild-type SINAT5 resulted in a reduced amount of NAC1 in roots, whereas overexpression of the C49S mutant led to higher quantities of NAC1 (Fig. 4c). In these transgenic plants, the levels of NAC1 transcript remained about the same. Treating the roots with auxin caused a reduction in the amount of NAC1 protein without any apparent effect on the amounts of the 35S-Myc-NAC1 transcript (Fig. 4d), which indicates that the hormone triggers post-translational proteolysis of the transcription activator.

Regulated proteolysis is important in the auxin signalling pathway^{6,7}. In the absence of auxin, the target genes for auxin signal transduction are presumably repressed by AUX/IAA transcriptional regulators. But on exposure to the hormone, these repressors are degraded by ubiquitin-dependent proteolysis mediated by the SCF complex, of which TIR1 is an integral subunit, leading to an upregulation of auxin signals. We have shown here that another step of the auxin signalling pathway is regulated by ubiquitin-mediated proteolysis. In this latter step, however, the ubiquitination is mediated not by an SCF complex but by the RING motif protein SINAT5, which shares extensive sequence homology with the SINA/

SIAH family of RING proteins^{10,11}. Much evidence supports the view that SINAT5 targets NAC1 for post-translational degradation: first, the amount of NAC1 protein *in vivo* can be increased by the proteasome inhibitor MG 132; second, NAC1 can be ubiquitinated by SINAT5 *in vitro*; third, NAC1 and SINAT5 are expressed in the same types of cell and interact *in vivo*; and last, overexpression of wild-type SINAT5 in transgenic plants leads to a decrease in the amount of NAC1 in roots and a concomitant reduction of lateral root numbers, whereas the opposite is found in transgenic plants overexpressing the dominant-negative C49S mutant of SINAT5. The opposing effects of SINAT5 and the C49S mutant are also observed when lateral root development is induced by auxin, thereby implicating the E3 ligase in this hormonal response.

Expression of both the SINAT5 and NAC1 genes is induced by auxin, although the former induction is slower. This difference in the induction kinetics could explain the auxin-induced degradation of NAC1, which becomes apparent at 8 h after hormone treatment (Figs 1f and 4d). The *Drosophila* protein SINA promotes degradation of the cell-fate repressor Tramtrack¹⁷, whereas mammalian SIAHs control the stability of the membrane receptor Deleted in Colorectal Cancer¹⁸, several transcription factors^{19,20} and other proteins^{21–24}. Our findings here provide insight into the function of the SINA/SIAH family as E3 ligases. The capacity of a SINAT5 protein containing a mutation in the RING motif to function as a



Figure 3 Phenotypes of transgenic plants. Seedlings of wild-type Landsberg ecotype (Control), transgenic lines carrying a 35S-SINAT5 transgene, SINAT5 (WT), and transgenic plants carrying a 35S-SINAT5(C49S) transgene, SINAT5 (C49S), were grown vertically on MS agar medium containing 2% sucrose. Representative plants were photographed after 10 d. Scale bar, 0.5 cm. The mean ($\pm$ s.d.) number of lateral roots per seedling and the mean length of main root per seedling (in parentheses) were determined for 30 seedlings of each type: control, 6.2 ± 0.3 (3.4 ± 0.4 cm); SINAT5 (WT), 2.7 ± 0.4 (4.6 ± 0.4 cm) and SINAT5 (C49S), 10.9 ± 0.2 (2.5 ± 0.3 cm).

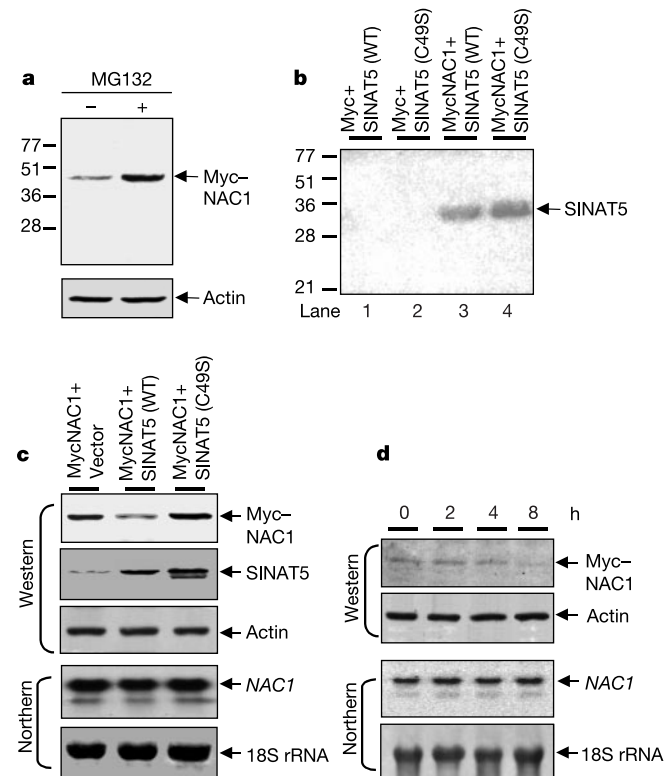


Figure 4 Relationship between amounts of SINAT5 and NAC1 proteins in transgenic plants. **a**, Expression of NAC1 is increased by treatment with MG 132. Root cultures from a Myc-NAC1 transgenic line were treated with 50 μ M MG 132 (+) or DMSO (–) for 3 h. Protein extracts were processed for western blot analysis. **b**, SINAT5 interacts with NAC1 *in vivo*. Myc-tagged proteins from the roots of double transgenic plants were immunoprecipitated and processed for western blot analysis. **c**, Relative amounts of SINAT5 and NAC1 in transgenic roots. Proteins from the roots of transgenic plants were processed for western blot analysis. Total RNAs from the same transgenic lines were analysed by northern blot. **d**, Auxin promotes the degradation of Myc-NAC1. Root cultures from a Myc-NAC1 transgenic line were treated with 2 μ M 1-NAA.

dominant-negative protein during *in vitro* ubiquitination suggests that the RING-independent dimerization of this protein family is functionally important for the E3 activity. Dimerization of RING-finger proteins as a means of activating E3 activity is in agreement with results obtained with heterodimers of BRAC1 and BARD1 (ref. 25).

In contrast to TIR1, which activates auxin signalling by protein degradation⁶, SINAT5 attenuates the signal by targeting NAC1 for degradation, thereby resetting the transduction cascade. The transcriptional activation of both *NAC1* (ref. 9) and *SINAT5* (Fig. 1e) by auxin indicates the importance of antagonistic transcriptional as well as post-transcriptional mechanisms in fine-tuning NAC1-regulated aspects of auxin signalling. At present we do not know whether other modifications of NAC1 precede its ubiquitination and how the self-ubiquitinating activity of SINAT5 and its ability to ubiquitinate NAC1 are regulated by auxin. The functions of four other SINAT family members in the *Arabidopsis* genome are also not known. Future investigations into these and other related molecular events will increase our understanding of the expanding role of ubiquitin-mediated proteolysis in cell signalling. □

Methods

Yeast two-hybrid screens

We synthesized complementary DNAs from 5 µg of poly(A) RNA isolated from plants at different growth stages using a commercial kit (Stratagene). The resulting double-stranded DNAs, with *EcoRI* and *XhoI* cut ends (average size 1.2 kilobases), were ligated to a pGAD-GH vector (Clontech) that had been digested with *EcoRI* and *XhoI* for 48 h at 8 °C. After ligation, the cDNA library was used to transform *E. coli* strain DH10B (Gibco) by electroporation. The cDNA library contains about 6 × 10⁶ primary transformants. We also isolated plasmids by standard methods. The yeast strain HF7c (*MATa ura3-52 his3-200 ade2-101 lys2-801 trp1-901 leu2-3, 112 gal4-542 gal80-538 LYS2::GAL1UAS-GAL1TATA-HIS3 URA3::GAL4 17mers* (× 3)–*CYC1TATA-LacZ*), containing the two reporter genes *lacZ* and *HIS3*, were used. Yeast cells were transformed sequentially with pGB-NAC1(1–199), a plasmid containing sequences encoding the first 199 amino acids of NAC1 fused to the Gal4 DNA-binding domain (a TRP1 marker) in the pGBT8 vector, and then with the *Arabidopsis* cDNA library constructed as described above. We carried out the screening as described²⁶.

Transgenic plants

A binary vector VIP96 (Kan^R in plants) carrying a 35S–6 × *Myc*–*NAC1* transgene was used to transform *A. thaliana* Landsberg ecotype. Wild-type plants and homozygous plants overexpressing *Myc*–*NAC1* were retransformed with pBA002 (Basta resistance in plants)²⁷ carrying either 35S–*SINAT5* or 35S–*SINAT5*(C49S). We generated all mutations using the Quick-Change site-directed mutagenesis kit (Stratagene). For the *SINAT5* promoter and *GUS* construct, we amplified by polymerase chain reaction (PCR) a 2-kilobase fragment upstream from the ATG codon of *SINAT5* and cloned it into pBI101 (Clontech). *Arabidopsis* plants were transformed by the floral dip method²⁸ and transformants were selected on MS plates containing 50 µg ml^{−1} kanamycin and/or 10 µg ml^{−1} Basta. The conditions for plant growth, induction of auxin and GUS staining have been described⁹. For the cellular localization of SINAT5, plasmid DNA containing a 35S–*SINAT5*–*GFP* construct was bombarded into onion peels as described⁹.

Protein preparation, *in vitro* binding and ubiquitination assays

We cloned DNA encoding SINAT5 and the SINAT5 (C49S) mutant into pMal-c2 (New England Biolabs) and prepared the fusion proteins according to the manufacturer's instructions. [³⁵S]methionine-labelled SINAT5 and NAC1 were generated by *in vitro* transcription and translation with wheat-germ extracts using a T7/T3 coupled TnT kit (Promega). For *in vitro* binding, 2 µl of the translation mix was added to 200 µl of binding buffer containing 50 mM HEPES (pH 7.5), 1 mM EDTA, 150 mM NaCl, 10% glycerol, 0.1% Tween 20 and 0.5 mM dithiothreitol (DTT), and the mixture was incubated at room temperature 25 °C for 1 h. After incubation, the beads containing amylose resin were washed three times with washing buffer (50 mM Tris-HCl (pH 7.5), 150 mM NaCl and 0.2% Nonidet P-40). ³²P-labelled ubiquitin was prepared using recombinant glutathione S-transferase (GST)-labelled ubiquitin as described^{14,16}. Unlabelled ubiquitin (10 µg per reaction) was purchased from Sigma. Preparation of wheat E1 (the cDNA clone was a kind gift from R. Vierstra) and UbcH5B (E2) has been described^{14,16}. For *in vitro* ubiquitination, MBP-labelled SINAT5 and/or the C49S mutant (500 ng) were immobilized on amylose resin beads. Ubiquitination assays were carried out by adding 20 ng each of recombinant wheat E1 and UbcH5B, and 2 × 10⁴ c.p.m. of ³²P-labelled ubiquitin in ubiquitination buffer containing 50 mM Tris (pH 7.4), 2 mM ATP, 5 mM MgCl₂, 2 mM DTT. The reaction mix (30 µl) was incubated for 1.5 h at 30 °C with agitation in an Eppendorf Thermomixer. After the reaction, samples were heated to 95 °C in an SDS-PAGE sample buffer containing β-mercaptoethanol before being separated by electrophoresis on 10% SDS gels.

Western and northern blot analyses

We raised antibodies to SINAT5 in rabbits using as an antigen a synthetic 14-residue

peptide (TDSIDSVIDDEIH) corresponding to amino acids 3–18 of SINAT5. Antibodies to c-Myc (SC40 and SC40-AC) and actin (SC1616) from Santa Cruz Biotechnology were used for immunoprecipitation and/or western blotting. Each lane contained 10 or 20 µg protein. We extracted RNA using an RNeasyR Plant Mini kit (Qiagen). Analysis by northern blot has been described⁹, using as probes an *NAC1* cDNA fragment encoding the NAC1 C-terminal region (amino acids 200–324) and a *SINAT5*-specific *EcoRI*/*NdeI* fragment. The 18S rRNA was used as a loading control. Each lane contained 10 µg RNA.

Received 13 February; accepted 25 June 2002; doi:10.1038/nature00998.

- del Pozo, J. C., Timpte, C., Tan, S., Callis, J. & Estelle, M. The ubiquitin-related protein RUB1 and auxin response in *Arabidopsis*. *Science* **280**, 1760–1763 (1998).
- Ruegger, M. *et al.* The TIR1 protein of *Arabidopsis* functions in auxin response and is related to human SKP2 and yeast grr1p. *Genes Dev.* **12**, 198–207 (1998).
- Rogg, L. E., Lasswell, J. & Bartel, B. A gain-of-function mutation in IAA28 suppresses lateral root development. *Plant Cell* **13**, 465–480 (2001).
- Ouellet, F., Overvoorde, P. J. & Theologis, A. IAA17/AXR3: biochemical insight into an auxin mutant phenotype. *Plant Cell* **13**, 829–841 (2001).
- Gray, W. M. *et al.* Identification of an SCF ubiquitin-ligase complex required for auxin response in *Arabidopsis thaliana*. *Genes Dev.* **13**, 1678–1691 (1999).
- Gray, W. M., Kepinski, S., Rouse, D., Leyser, O. & Estelle, M. Auxin regulates SCF(TIR1)-dependent degradation of AUX/IAA proteins. *Nature* **414**, 271–276 (2001).
- Ramos, J. A., Zenser, N., Leyser, O. & Callis, J. Rapid degradation of auxin/indoleacetic acid proteins requires conserved amino acids of domain II and is proteasome dependent. *Plant Cell* **13**, 2349–2360 (2001).
- Zenser, N., Ellsmore, A., Leasure, C. & Callis, J. Auxin modulates the degradation rate of Aux/IAA proteins. *Proc. Natl Acad. Sci. USA* **98**, 11795–11800 (2001).
- Xie, Q., Frugis, G., Colgan, D. & Chua, N. H. *Arabidopsis* NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. *Genes Dev.* **14**, 3024–3036 (2000).
- Carthew, R. W. & Rubin, G. M. Seven in absentia, a gene required for specification of R7 cell fate in the *Drosophila* eye. *Cell* **63**, 561–577 (1990).
- Nemani, M. *et al.* Activation of the human homologue of the *Drosophila* sina gene in apoptosis and tumour suppression. *Proc. Natl Acad. Sci. USA* **93**, 9039–9342 (1996).
- Hu, G. & Fearon, E. R. Siah-1 N-terminal RING domain is required for proteolysis function, and C-terminal sequences regulate oligomerization and binding to target proteins. *Mol. Cell. Biol.* **19**, 724–732 (1999).
- Polekhina, G. *et al.* Siah ubiquitin ligase is structurally related to TRAF and modulates TNF-α signaling. *Nature Struct. Biol.* **9**, 68–75 (2002).
- Lorick, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl Acad. Sci. USA* **96**, 11364–11369 (1999).
- Joazeiro, C. A. *et al.* The tyrosine kinase negative regulator c-Cbl as a RING-type, E2 dependent ubiquitin-protein ligase. *Science* **286**, 309–312 (1999).
- Fang, S., Jensen, J. P., Ludwig, R. L., Vousden, K. H. & Weissman, A. M. Mdm2 is a RING finger-dependent ubiquitin protein ligase for itself and p53. *J. Biol. Chem.* **275**, 8945–8951 (2000).
- Li, S., Li, Y., Carthew, R. W. & Lai, Z. C. Photoreceptor cell differentiation requires regulated proteolysis of the transcriptional repressor Tramtrack. *Cell* **90**, 469–478 (1997).
- Hu, G. *et al.* Mammalian homologs of seven in absentia regulate DCC via the ubiquitin-proteasome pathway. *Genes Dev.* **11**, 2701–2747 (1997).
- Tiedt, R., Bartholdy, B. A., Matthias, G., Newell, J. W. & Matthias, P. The RING finger protein Siah-1 regulates the level of the transcriptional coactivator OBF-1. *EMBO J.* **20**, 4143–4152 (2001).
- Boehm, J., He, Y., Greiner, A., Staudt, L. & Wirth, T. Regulation of BOB.1/OBF.1 stability by SIAH. *EMBO J.* **20**, 4153–4162 (2001).
- Nie, J. *et al.* LNX functions as a RING type E3 ubiquitin ligase that targets the cell fate determinant Numb for ubiquitin-dependent degradation. *EMBO J.* **21**, 93–102 (2002).
- Susini, L. *et al.* Siah-1 binds and regulates the function of Numb. *Proc. Natl Acad. Sci. USA* **98**, 15067–15072 (2001).
- Matsuzawa, S. & Reed, J. C. Siah-1 SIP and Ebi collaborate in a novel pathway for β-catenin degradation linked to p53 responses. *Mol. Cell* **7**, 915–926 (2001).
- Liu, J. *et al.* Siah-1 mediates a novel β-catenin degradation pathway linking p53 to the adenomatous polyposis coli protein. *Mol. Cell* **7**, 927–936 (2001).
- Hashizume, R. *et al.* A RING heterodimer BRCA1–BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. *J. Biol. Chem.* **276**, 14537–14540 (2001).
- Xie, Q., Sanz-Burgos, A. P., Guo, H., Garcia, J. A. & Gutierrez, C. GRAB proteins, novel members of the NAC domain family, isolated by their interaction with a geminivirus protein. *Plant Mol. Biol.* **39**, 647–656 (1999).
- Kost, B., Spielhofer, P. & Chua, N. H. A GFP-mouse talin fusion protein labels plant actin filaments *in vivo* and visualizes the actin cytoskeleton in growing pollen tubes. *Plant J.* **16**, 393–401 (1998).
- Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).

Acknowledgements

We thank E. Ng and W.-P. Tang for technical assistance; Y.-S. Chan for taking the GUS pictures; and P. Hare for reading the manuscript. This work was supported in part by a grant from the NIH to N.-H.C. This work was initiated in the Institute of Molecular Agrobiology, Singapore, supported by Singapore A-star funding.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.-H.C. (e-mail: chua@mail.rockefeller.edu).

L23 protein functions as a chaperone docking site on the ribosome

Günter Kramer^{*,†}, Thomas Rauch^{*,†}, Wolfgang Rist[‡],
Sonja Vorderwülbecke[‡], Holger Patzelt[‡], Agnes Schulze-Specking[‡],
Nenad Ban[§], Elke Deuerling[‡] & Bernd Bukau[‡]

^{*} Institut für Biochemie und Molekularbiologie, Universität Freiburg, Hermann-Herder-Strasse 7, 79104 Freiburg, Germany

[‡] Zentrum für Molekulare Biologie (ZMBH), Universität Heidelberg, INF282, D-69120 Heidelberg, Germany

[§] Institute for Molecular Biology and Biophysics, Swiss Federal Institute of Technology, ETH Hönggerberg, HPK Building, CH-8093 Zürich, Switzerland

[†] These authors contributed equally to this work

During translation, the first encounter of nascent polypeptides is with the ribosome-associated chaperones that assist the folding process—a principle that seems to be conserved in evolution^{1–3}. In *Escherichia coli*, the ribosome-bound Trigger Factor chaperones the folding of cytosolic proteins by interacting with nascent polypeptides^{4,5}. Here we identify a ribosome-binding motif in the amino-terminal domain of Trigger Factor. We also show the formation of crosslinked products between Trigger Factor and two adjacent ribosomal proteins, L23 and L29, which are located at the exit of the peptide tunnel in the ribosome. L23 is essential for the growth of *E. coli* and the association of Trigger Factor with the ribosome, whereas L29 is dispensable in both processes. Mutation of an exposed glutamate in L23 prevents Trigger Factor from interacting with ribosomes and nascent chains, and causes protein aggregation and conditional lethality in cells that lack the protein repair function of the DnaK chaperone. Purified L23 also interacts specifically with Trigger Factor *in vitro*. We conclude that essential L23 provides a chaperone docking site on ribosomes that directly links protein biosynthesis with chaperone-assisted protein folding.

The N-terminal domain of Trigger Factor (TF) mediates its binding to the 50S ribosomal subunit⁶. We aligned the N-terminal domains of TF homologues and found a conserved sequence of about 17 residues, including a completely conserved Gly-Phe-Arg-x-Gly-x-x-Pro motif, called the TF signature (Fig. 1a), which is predicted to be localized in an unstructured region. This prediction, and our observation that this stretch of sequence is susceptible to proteolysis, indicated that it could be exposed on the surface and thus might contribute to the interaction of TF with ribosomes.

To investigate whether this TF signature is involved in ribosome binding, we replaced residues Phe 44, Arg 45 and Lys 46 with alanine to form an FRK/AAA mutant of TF (Fig. 1a). Circular dichroism, partial proteinase K digestion and gel filtration verified the structural integrity of the FRK/AAA protein. We then tested whether it could associate with ribosomes purified from *Δtig* (which encodes TF) *E. coli* cells. After incubating the ribosomes with either TF or the FRK/AAA mutant, ribosome-TF complexes were separated from unbound TF by centrifugation. The FRK/AAA mutant protein was reduced in its association with ribosomes (Fig. 1b, lanes 3 and 5), indicating that the TF signature is involved in the interaction of TF with ribosomes.

To crosslink TF to its ribosomal binding site, we first replaced Asp 42, which is adjacent to the signature motif, with cysteine (TF D42C, Fig. 1a). The lack of additional cysteines in TF then allowed us to couple the thiol-specific ultraviolet-activatable crosslinker benzophenone-4-iodoacetamide (BPIA) to Cys 42. Because BPIA preferentially attacks C–H bonds, it can crosslink TF to ribosomal proteins and ribosomal RNA. The TF D42C mutant was not altered in structural integrity or in ribosome binding (data not shown).

After being coupled to BPIA, the TF D42C mutant was incubated

with ribosomes from *Δtig* cells. Ultraviolet irradiation resulted in two additional bands with relative molecular masses (M_r) of 68,000 (68K) and 75K in the ribosomal pellet. These bands represented crosslinking products between TF and ribosomes because they depended strictly on the presence of ribosomes and were absent in the supernatant (Fig. 1b, lanes 6 and 7) and in non-irradiated samples (data not shown). RNase A treatment did not alter the mobility of these products, which indicated that ribosomal RNA was not present. The crosslinking bands were excised and digested with trypsin, and the resulting peptides were sequenced by nano-electrospray mass spectrometry (Supplementary Information Fig. 1a, b). We detected TF-derived peptides in both crosslinking bands. Unique peptides originating from ribosomal proteins L29 and L23 were identified in the 68K and 75K crosslinking products, respectively. BPIA is flexible and has a length of about 10 Å, which explains how it can form crosslinks to two adjacent proteins (see below).

The possibility that the TF D42C–BPIA mutant associated non-specifically with ribosomes was excluded by a competition experiment. We simultaneously added TF D42C–BPIA and a 2.5-fold molar excess of TF to ribosomes and analysed the crosslinking efficiency. Whereas wild-type TF competed with TF D42C–BPIA for ribosome binding and decreased the occurrence of both crosslinking products, TF FRK/AAA did not (data not shown). The

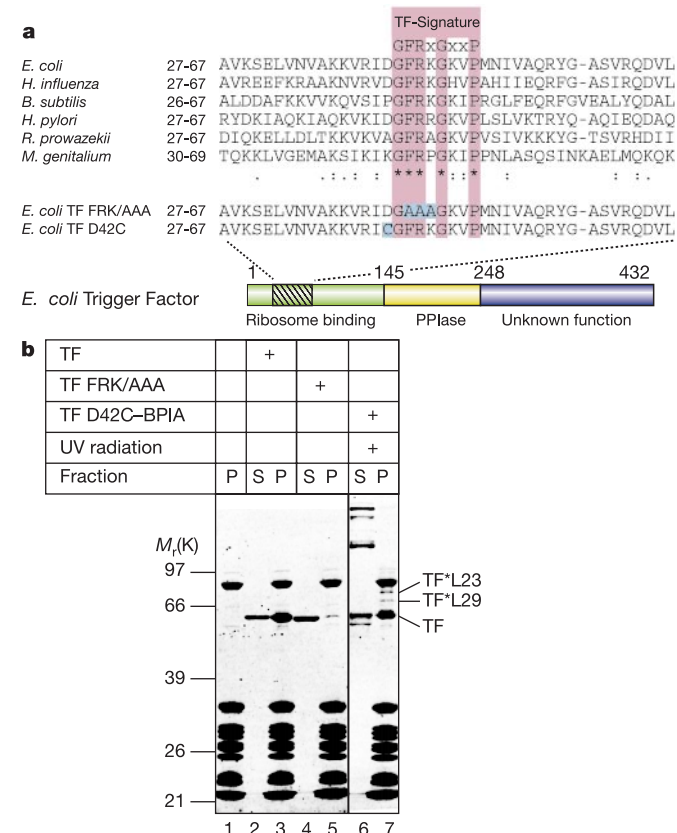


Figure 1 Trigger Factor is crosslinked to the ribosomal proteins L23 and L29. **a**, Alignment of the N-terminal regions of bacterial TF homologues using CLUSTALW¹⁸, and the domain structure of *E. coli* TF. Identical amino acid residues are marked with an asterisk and shown in purple; conserved and semi-conserved residues are indicated by colons and dots, respectively. Amino acid residues that were replaced in TF mutant proteins are boxed in blue. **b**, Ribosomes from *Δtig* cells were incubated with TF variants followed by sucrose cushion centrifugation to separate the ribosomal pellets (P) from the soluble proteins (S). Fractions were separated by SDS–PAGE and stained with Coomassie blue. Where indicated, samples were irradiated with ultraviolet.

crosslinking products obtained with TF D42C–BPIA thus resulted from a specific interaction of TF with ribosomes.

There are homologues of L23 and L29 in all kingdoms. The crystal structure of the 50S subunit from *Haloarcula marismortui* shows that L23 and L29 are in direct contact with each other and are located next to the exit of the tunnel for nascent polypeptides^{7,8}. Although *H. marismortui* lacks TF, the TF D42C–BPIA mutant crosslinked specifically to L23 of the *H. marismortui* 50S ribosomal subunits (refs 7, 8, and unpublished data), which ruled out the

possibility that additional non-ribosomal or ribosome-associated factors are required for TF to bind to ribosomes.

To determine whether TF associates directly with L23 and/or L29, we generated *E. coli* strains with mutations in the genes *rplW* and *rpmC*, which encode L23 and L29, respectively. We reasoned that if there is a direct interaction between TF and L23 or L29, then we would observe reduced binding of TF to ribosomes originating from such strains. We first attempted to delete the *rplW* and *rpmC* genes in MC4100 by replacing the corresponding open reading frames with a nonpolar kanamycin resistance cassette⁹. MC4100 $\Delta rplW::kan$ mutant cells were viable at all temperatures tested (23–37 °C), although they grew slightly slower than did wild-type strains on rich media plates (Fig. 2a). L29 is thus dispensable for the growth and protein biosynthesis of *E. coli*.

To test whether the absence of L29 influences TF binding, we purified ribosomes from $\Delta rpmC::kan$ mutant cells in the presence of a high salt concentration, which minimizes unspecific associations of proteins with ribosomes but allows the reduced but detectable binding of TF¹⁰. Similar amounts of TF were associated with ribosomes prepared from $\Delta rpmC::kan$ mutant and equivalent wild-type cells (Fig. 2b, lanes 4 and 12). We then analysed whether TF could rebind to mutant ribosomes under physiological salt conditions. Ribosomes purified in high salt were incubated with a twofold molar excess of TF in low-salt buffer followed by centrifugation. Ribosomes lacking L29 were not impaired in their association with TF (Fig. 2c, lanes 4 and 12), showing that L29 does not contribute significantly to TF binding to ribosomes.

Attempts to delete the chromosomal *rplW* gene failed unless the gene was expressed from a plasmid (pL23). Growth of MC4100 $\Delta rplW::kan$ cells containing pL23 was dependent on isopropylthiogalactoside (IPTG), which induced the expression of plasmid-encoded *rplW* (Fig. 2a). We designed L23 variants with alterations in residues whose side chains were judged to be exposed at the ribosomal surface (from the structure of the *H. marismortui* ribosome), and that are conserved among eubacterial L23 proteins. We identified two regions (1 and 2) in L23 that contain residues fulfilling both criteria (Supplementary Information Fig. 1c). The plasmid-encoded *rplW* gene was mutated to introduce alterations in region 1 (Glu 18 to alanine, yielding L23 E18A; Glu 18 to glutamine, yielding L23 E18Q; and Val 16, Ser 17 and Glu 18 to alanine, yielding L23 VSE/AAA) and region 2 (Glu 52 to lysine, yielding L23 E52K; Phe 51, Glu 52 and Val 53 to alanine, yielding L23 FEV/AAA). Each mutant *rplW* allele complemented the lethal phenotype of $\Delta rplW::kan$ cells at 30 °C and 37 °C (the results for L23 E18A are shown in Fig. 2a), which allowed us to isolate ribosomes containing mutant L23 proteins.

Ribosomes from MC4100 $\Delta rplW::kan$ cells that expressed L23 variants were tested for their co-purification with TF under physiological salt concentrations (Fig. 2d). Whereas alterations in region 2 did not influence TF association with ribosomes, all alterations in region 1 (E18A, E18Q and VSE/AAA) strongly reduced TF binding and resulted in a reduced co-purification of TF with ribosomes under high salt conditions (Fig. 2b). The rebinding of TF to these ribosomes under physiological salt conditions was also severely decreased (Fig. 2c, compare lanes 4, 6, 8 and 10). All ribosome preparations contained comparable amounts of L23 and S1 (Fig. 2c), which excluded the possibility that a smaller incorporation of mutant L23 proteins into ribosomes was responsible for the reduced binding.

To verify that TF interacted directly with L23, we purified an S-tagged L23 variant fused to thioredoxin to increase its solubility and coupled the protein (Trx–L23) to S-tag agarose columns. TF or the FRK/AAA mutant was applied to the columns, and bound proteins were eluted (Fig. 2e). Whereas TF bound to the L23 fusion protein, the FRK/AAA mutant showed a substantial decrease in its association with L23. Because TF and the FRK/AAA mutant have similar substrate-binding properties *in vitro* (unpublished data), we could

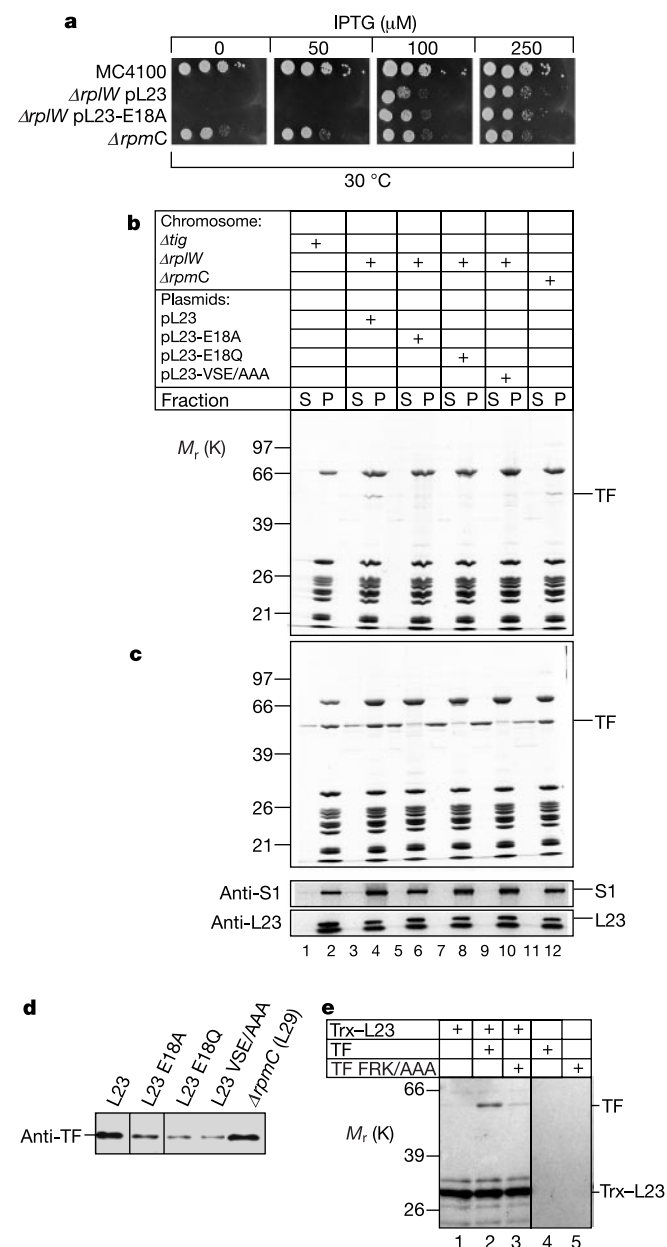


Figure 2 Ribosomes with mutations in L23 are deficient in TF binding. **a**, Growth analysis. **b**, Mutant ribosomes purified in high salt were analysed for TF association by SDS–PAGE and Coomassie blue staining. Δtig ribosomes are shown as control. **c**, Ribosomes (2 μ M) were incubated with TF (4 μ M) under physiological salt conditions. Ribosomal pellets (P) and supernatants (S) were analysed as described in **b**. L23 and S1 contents were analysed by western blotting. The lower signal in the anti-L23 blot is an unspecific crossreaction. **d**, Mutant ribosomes were purified under physiological salt conditions and TF content was analysed by immunoblotting. **e**, TF or the TF FRK/AAA mutant was incubated with Trx–L23 bound to S-tag agarose. Eluates were analysed by SDS–PAGE and Coomassie blue staining.

exclude the possibility of a chaperone–substrate interaction of TF and Trx–L23. Thus, L23 interacts directly with TF on ribosomes.

To investigate whether ribosome binding is a requirement for the interaction of TF with nascent polypeptides, we generated an *in vitro* transcription/translation system comprising a translation-competent (ribosome-free) fraction prepared from MC4100 Δ tig cells¹¹, purified TF, and ribosomes containing either wild-type or mutant L23 proteins. Ribosome–nascent chain complexes (RNCs) carrying the ³⁵S-labelled N-terminal fragment (residues 1–173) of *E. coli* isocitrate dehydrogenase (ICDH), which is an *in vivo* substrate of TF¹², were produced. TF association with arrested nascent ICDH was analysed by crosslinking using the amino-reactive homobifunctional agent disuccinimidyl suberate (DSS). RNCs containing wild-type L23 produced two prominent crosslinking products (Fig. 3a). Both products were co-immunoprecipitated with antibodies specific for TF, which thus indicated that crosslinks had formed between TF and nascent polypeptides. Although observed previously with different RNCs¹², the reasons why there

are two crosslinking products are unknown. The same products with comparable intensity were formed with L29-deficient RNCs (Fig. 3a, lanes 1–3 and 13–15). By contrast, ribosomes containing L23 mutant proteins with alterations at residue Glu 18 (E18A, E18Q and VSE/AAA) severely reduced TF interaction with nascent ICDH (Fig. 3a, lanes 4–12). The binding of TF to L23 through Glu 18 is therefore a prerequisite for the association of TF with nascent chains.

We next investigated whether the *in vivo* role of TF in protein folding requires its association with ribosomes. We took into account the fact that TF and the DnaK system cooperate in protein folding, which means that Δ tig Δ dnaK mutants are synthetically lethal at 37 °C (refs 4, 5). We tested whether a Δ rplW::kan deletion could be introduced into Δ dnaK52 cells that express plasmid-encoded wild-type or mutant L23 proteins. The rationale was that if TF association with ribosomes is a prerequisite for its function *in vivo*, then it should be possible to detect growth defects and/or protein aggregation in Δ dnaK52 cells whose ribosomes contain L23 mutant proteins. The Δ rplW::kan allele could be introduced into Δ dnaK52 mutants expressing wild-type or mutant L23 proteins without growth defects at 30 °C in Luria broth. At 34 °C, however, the Δ rplW::kan Δ dnaK52 cells expressing the E18A, E18Q and VSE/AAA variants of L23 grew more slowly than did those expressing wild-type L23, and at 37 °C the expression of all three L23 mutant alleles was synthetically lethal (Fig. 3b).

To investigate the extent of protein misfolding in these cells, we grew Δ rplW::kan Δ dnaK52 cells expressing L23 mutants from plasmids for four doubling times at 32 °C, 33 °C and 37 °C. These cells could grow for four doubling times at 37 °C in liquid media, perhaps because the aggregated protein did not reach a critical amount in this short period. In cells expressing L23 mutant proteins, the amount of insoluble protein aggregates increased markedly with rising temperature and was most severe in cells expressing the VSE/AAA mutant of L23 (Fig. 3c). In Δ dnaK52 Δ rplW cells complemented with L23 mutant proteins, the amount of ribosome-bound TF was thus too low to prevent protein aggregation at elevated temperatures. This is in agreement with our finding that strong overproduction of the FRK/AAA mutant complements lethality of Δ dnaK Δ tig only up to 32 °C (data not shown).

We conclude that L23 is the docking site for TF and thereby

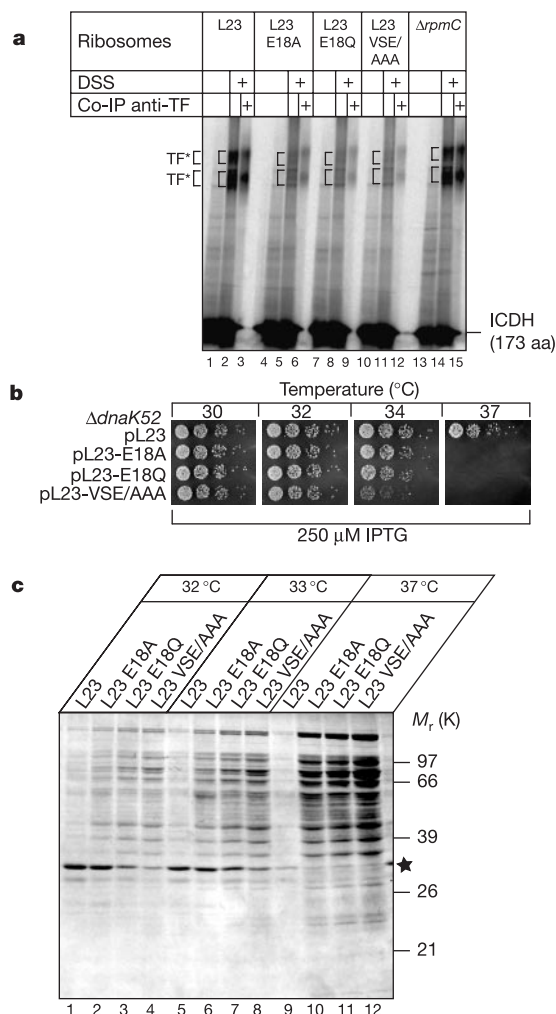


Figure 3 L23 docking is essential for the *in vivo* function of TF. **a**, L23 point mutations influence the association of TF with nascent polypeptides. Arrested ³⁵S-labelled nascent ICDH was synthesized *in vitro* using different ribosomes. After crosslinking and centrifugation, RNCs were co-immunoprecipitated to identify TF crosslinks. Brackets indicate TF crosslinks. aa, amino acids. **b**, Temperature sensitivity of Δ dnaK52 Δ rplW cells expressing different L23 alleles. **c**, Δ dnaK52 Δ rplW cells expressing different rplW alleles were collected in log phase. Aggregates were isolated¹⁹ and analysed by SDS-PAGE and Coomassie blue staining. The band marked by an asterisk probably corresponds to an outer membrane protein. The reason for its disappearance at 37 °C is unknown.

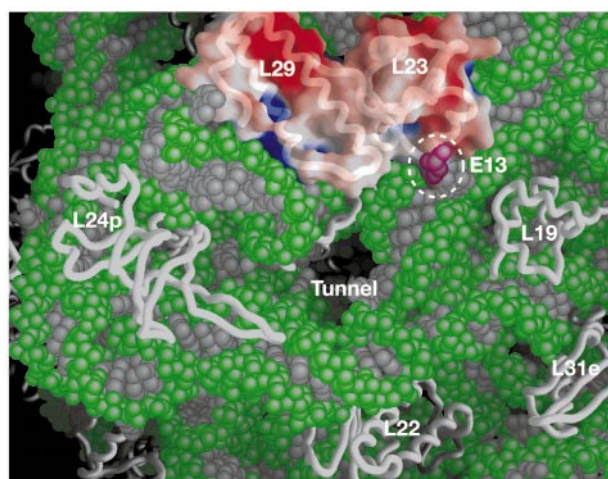


Figure 4 Positioning of the conserved glutamate in L23 at the polypeptide exit tunnel. Space-filling model of the ribosomal polypeptide exit region of the 50S subunit from *H. marismortui*. L23 and L29 are shown in a surface-charge distribution illustration²⁰ generated by GRASP. The surface-exposed Glu 13 (corresponding to *E. coli* Glu 18) residue that is involved in the interaction with TF is shown in purple. Other proteins in the exit region are visible by their C α traces.

couples protein biosynthesis with chaperone-assisted protein folding. The involvement of Glu 18 in L23 (which corresponds to Glu 13 in *H. marismortui* L23)—a residue that is located at the rim of the peptide tunnel—in the binding process indicates that TF is positioned exactly where nascent polypeptides emerge from the tunnel (Fig. 4). This positioning is essential for the chaperone activity of TF. □

Methods

Strains and genetic manipulations

All *E. coli* strains were derivatives of MC4100. We grew cells in Luria broth medium supplemented with 100 µg ml⁻¹ ampicillin, 40 µg ml⁻¹ kanamycin and 250 µM IPTG when appropriate. The methods for P1 lysates, P1 transductions and disrupting chromosomal genes have been described^{21,3}. Plasmids pL23, pHL23 and pL29 are derivatives of pTrc99B (ref. 14), and plasmid pFTR6 (encoding Trx–L23) is a derivative of pET32a (Novagen). We amplified *rplW* and *rpmC* by polymerase chain reaction using MC4100 chromosomal DNA as a template and cloned the products individually into pTrc99B to generate pL23 and pL29, respectively, and into pTrc-6His (lab collection) to generate pHL23 for expressing His₆–L23. Mutations in *rplW* were created using the QuikChange Site-Directed Mutagenesis Kit (Stratagene) and pL23 as the template.

Protein purification and antibody production

We purified TF and TF fragments as described⁶. To raise antibodies against L23, His₆–L23 was expressed from plasmid pHL23 and purified under denaturing conditions using Ni²⁺-NTA (Qiagen).

In vitro binding analysis

The Trx–L23 fusion protein was expressed from pFTR6 vector and purified using Ni²⁺-NTA (Qiagen). S-tag agarose (Novagen) was saturated with purified Trx–L23 and unbound protein was removed by intense washing. We applied TF or the TF FRK/AAA mutant to the column and incubated it for 30 min at 23 °C. After washing with ten column volumes of buffer, bound proteins were eluted with 0.2 M sodium citrate, pH 2.

Crosslinking

We coupled BPIA (Molecular Probes) to the D42C mutant of TF protein as described¹⁵. Crosslinking mixtures were incubated for 30 min at 30 °C and subsequently irradiated on ice for 5 min with ultraviolet (365 nm, 100 W; Model B-100AP, Ultraviolet Products) at a distance of 5 cm. We separated ribosomal complexes by centrifugation through sucrose cushions as described⁶.

Protein identification by mass spectrometry

Bands were excised from one-dimensional SDS–PAGE gels stained with Coomassie blue and digested in gel with trypsin as described¹⁶. We analysed the tryptic peptides by nanoelectrospray tandem mass spectrometry as described¹⁷ using a QSTAR Pulsar (MDS Sciex) equipped with a nanoES ion source (MDS Proteomics). Sequence searches were done with the Protein and Peptide Software Suite (MDS Proteomics) in a non-redundant database.

In vitro transcription/translation

We carried out transcription and translation assays as described¹¹.

Received 18 June; accepted 11 July 2002; doi:10.1038/nature01047.

- Hartl, F. U. & Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).
- Bukau, B., Deuerling, E., Pfund, C. & Craig, E. A. Getting newly synthesized proteins into shape. *Cell* **101**, 119–122 (2000).
- Frydman, J. Folding of newly translated proteins *in vivo*: the role of molecular chaperones. *Annu. Rev. Biochem.* **70**, 603–647 (2001).
- Deuerling, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A. & Bukau, B. Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature* **400**, 693–696 (1999).
- Teter, S. A. *et al.* Polypeptide flux through bacterial Hsp70: DnaK cooperates with Trigger Factor in chaperoning nascent chains. *Cell* **97**, 755–765 (1999).
- Hesterkamp, T., Deuerling, E. & Bukau, B. The amino-terminal 118 amino acids of *Escherichia coli* trigger factor constitute a domain that is necessary and sufficient for binding to ribosomes. *J. Biol. Chem.* **272**, 21865–21871 (1997).
- Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
- Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
- Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
- Hesterkamp, T., Hauser, S., Lütcke, H. & Bukau, B. *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proc. Natl Acad. Sci. USA* **93**, 4437–4441 (1996).
- Schaffitzel, E., Rüdiger, S., Bukau, B. & Deuerling, E. Functional dissection of Trigger Factor and DnaK: Interactions with nascent polypeptides and thermally denatured proteins. *J. Biol. Chem.* **382**, 1235–1243 (2001).
- Patzelt, H. *et al.* Binding specificity of *Escherichia coli* trigger factor. *Proc. Natl Acad. Sci. USA* **98**, 14244–14249 (2001).
- Miller, J. H. *Experiments in Molecular Genetics* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1972).

- Amann, E., Ochs, B. & Abel, K. J. Tightly regulated *tac* promoter vectors useful for the expression of unfused and fused proteins in *Escherichia coli*. *Gene* **69**, 301–315 (1988).
- Laufen, T. *et al.* Mechanism of regulation of Hsp70 chaperones by DnaJ co-chaperones. *Proc. Natl Acad. Sci. USA* **96**, 5452–5457 (1999).
- Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).
- Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466–469 (1996).
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
- Tomoyasu, T., Mogk, A., Langen, H., Goloubinoff, P. & Bukau, B. Genetic dissection of the roles of chaperones and proteases in protein folding and degradation in the *Escherichia coli* cytosol. *Mol. Microbiol.* **40**, 397–413 (2001).
- Nicholls, A., Sharp, K. A. & Honig, B. J. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Protein Struct. Funct. Genet.* **11**, 281–296 (1991).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank the members of the Bukau lab for discussions; T. Hesterkamp for suggesting the TF FRK/AAA mutation; K. Turgay for suggesting the S-tag experiment; and D. Dougan for comments on the manuscript. This work was supported by grants of the Deutsche Forschungsgemeinschaft to B.B. and E.D., the Human Frontier Science Program to E.D. and N.B., the Swiss National Science Foundation to N.B., and fellowships of the Boehringer Ingelheim Fonds to T.R. and the Fonds der Chemischen Industrie (Kékulé scholarship) to W.R.

Correspondence and requests should be addressed to B.B.
(e-mail: bukau@zmbh.uni-heidelberg.de) or E.D.
(e-mail: E.Deuerling@zmbh.uni-heidelberg.de).

Oxidative demethylation by *Escherichia coli* AlkB directly reverts DNA base damage

Sarah C. Treweek^{*,†}, Timothy F. Henshaw^{†,‡}, Robert P. Hausinger[‡], Tomas Lindahl^{*} & Barbara Sedgwick^{*}

^{*} Cancer Research UK London Research Institute, Clare Hall Laboratories, South Mimms, Hertfordshire EN6 3LD, UK

[‡] Departments of Chemistry or Microbiology & Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA

[†] These authors contributed equally to this work

Methylating agents generate cytotoxic and mutagenic DNA damage. Cells use 3-methyladenine-DNA glycosylases to excise some methylated bases from DNA, and suicidal O⁶-methylguanine-DNA methyltransferases to transfer alkyl groups from other lesions onto a cysteine residue^{1,2}. Here we report that the highly conserved AlkB protein repairs DNA alkylation damage by means of an unprecedented mechanism. AlkB has no detectable nuclease, DNA glycosylase or methyltransferase activity; however, *Escherichia coli* *alkB* mutants are defective in processing methylation damage generated in single-stranded DNA^{3–5}. Theoretical protein fold recognition had suggested that AlkB resembles the Fe(II)- and α -ketoglutarate-dependent dioxygenases⁶, which use iron-oxo intermediates to oxidize chemically inert compounds^{7,8}. We show here that purified AlkB repairs the cytotoxic lesions 1-methyladenine and 3-methylcytosine in single- and double-stranded DNA in a reaction that is dependent on oxygen, α -ketoglutarate and Fe(II). The AlkB enzyme couples oxidative decarboxylation of α -ketoglutarate to the hydroxylation of these methylated bases in DNA, resulting in direct reversion to the unmodified base and the release of formaldehyde.

DNA alkylating agents occur endogenously, are present in the environment and are used in chemotherapy^{9,10}. *Escherichia coli* exposed to these compounds respond by inducing the expression of four genes, *ada*, *alkA*, *aidB* and *alkB*. Ada protein is an O⁶-methylguanine-DNA methyltransferase and also regulates this adaptive response. AlkA is a 3-methyladenine-DNA glycosylase, and AidB is proposed to destroy certain alkylating agents^{1,2}. AlkB is conserved from bacteria to mammals, but its role has not been resolved despite the early isolation of an *E. coli* *alkB* mutant³. Expression of *E. coli* *alkB* confers alkylation resistance to human cells¹¹, and conversely, a human homologue has been reported to convey methyl methanesulphonate (MMS) resistance to the *E. coli* mutant¹². AlkB processes the cytotoxic DNA damage generated in single-stranded DNA by S_N2 methylating agents, such as MMS, dimethylsulphate (DMS) and methyl iodide⁵. 1-Methyladenine and 3-methylcytosine are predominant forms of base damage only in single-stranded DNA because the modification sites are normally protected by base pairing^{13,14}. These lesions are cytotoxic because they stall DNA replication, and are not removed by known DNA repair pathways. We previously proposed 1-methyladenine and 3-methylcytosine in DNA as candidate substrates of the AlkB protein⁵.

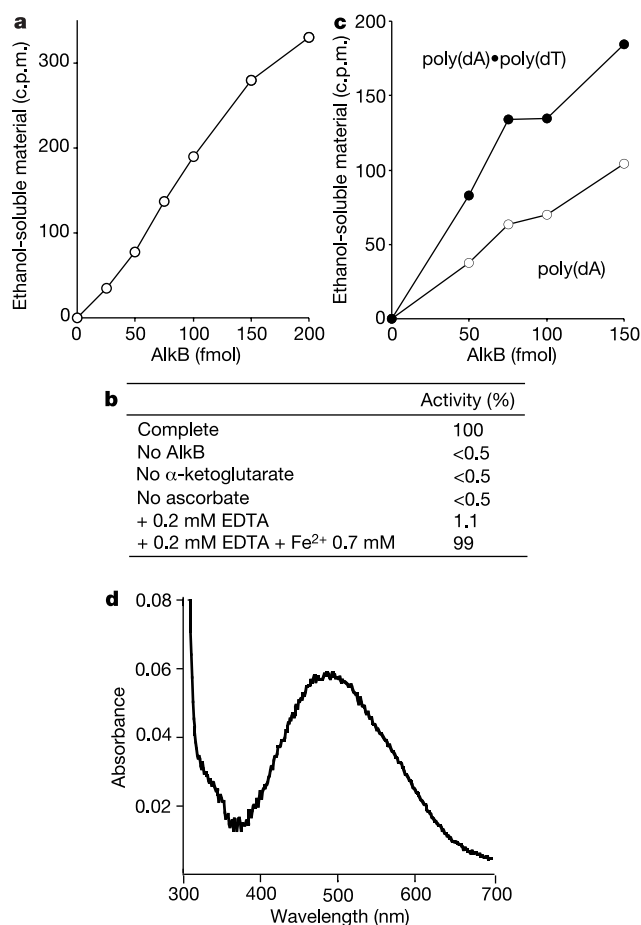


Figure 1 Release of ethanol-soluble material from methylated poly(dA) by AlkB in the presence of Fe(II) and α-ketoglutarate. **a**, [¹⁴C]Methyl iodide-treated poly(dA) (1,200 c.p.m.) was incubated with AlkB in the complete reaction mixture, and the release of radioactive material was monitored. **b**, Requirements for AlkB activity (2 pmol AlkB). **c**, Comparison of AlkB activity on [¹⁴C]-labelled methylated poly(dA) (1,200 c.p.m.) (open circles) and after annealing to poly(dT) (filled circles), assayed at 20 °C for 30 min to maintain stable duplexes. **d**, Difference absorption spectrum of anaerobic AlkB (0.35 mM) in the presence of α-ketoglutarate (1 mM) and Fe(II) (0.3 mM) minus the spectrum without Fe(II).

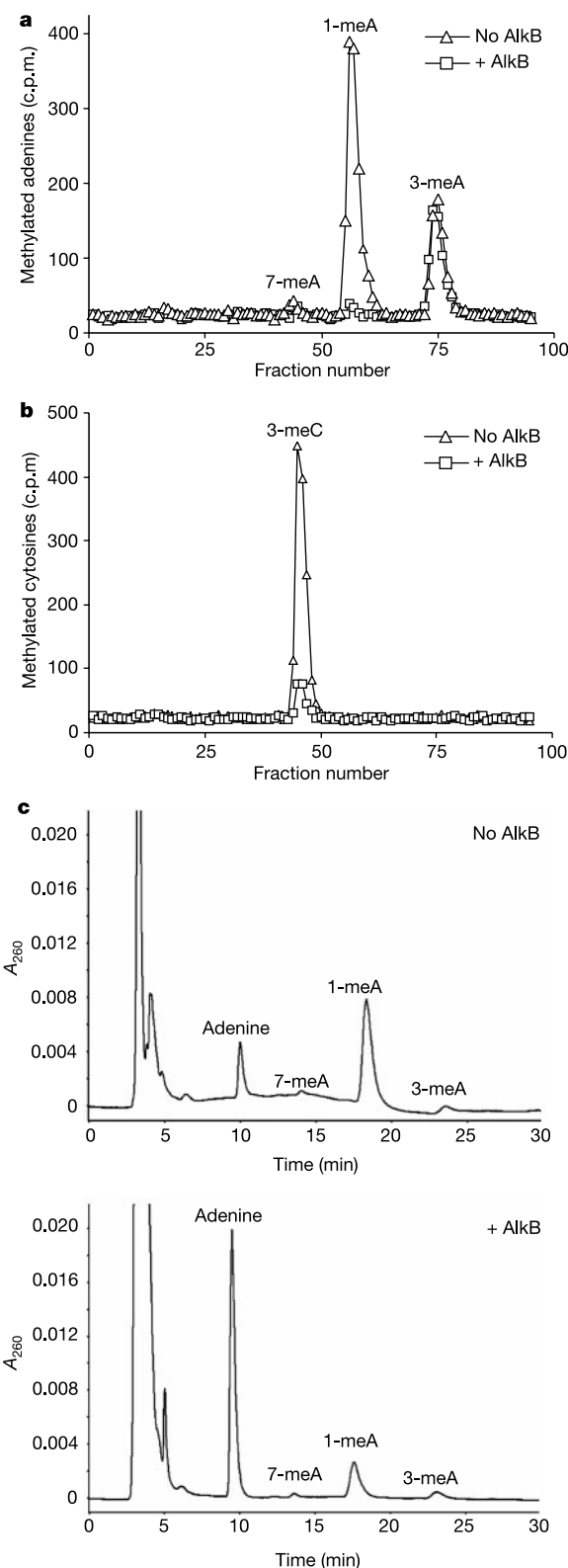


Figure 2 Repair of 1-methyladenine (1-meA) and 3-methylcytosine (3-meC) by AlkB. **a**, **b**, [¹⁴C]Methyl iodide-treated poly(dA) (**a**) or poly(dC) (**b**) were incubated without AlkB (triangles) or with 2.5 pmol AlkB (squares). The [¹⁴C]-labelled methylated bases remaining in the substrates were analysed by HPLC and scintillation counting. 3-meA and 7-meA, 3- and 7-methyladenine, respectively. **c**, Direct reversion of 1-methyladenine to adenine by AlkB. A thymine-rich oligodeoxynucleotide containing five adenine residues was heavily methylated and then incubated without (top) or with (bottom) 920 pmol AlkB at 37 °C for 30 min. The bases present in the oligonucleotide were analysed by HPLC and A₂₆₀ measurements. The early eluting peaks were oligo(dT) fragments.

Nevertheless, many attempts to develop assays for this enzyme were unsuccessful. Theoretical sequence profile and fold recognition searches suggested that AlkB may be either a hydrolase or an α -ketoglutarate- and Fe(II)-dependent dioxygenase^{6,15}. Here we demonstrate DNA repair activity for AlkB using DNA substrates containing 1-methyladenine or 3-methylcytosine, α -ketoglutarate as a co-substrate, and Fe(II) as a cofactor.

A substrate containing ¹⁴C-labelled methylated adenine residues was prepared by treating poly(dA) with [¹⁴C]methyl iodide. In the presence of α -ketoglutarate and Fe(II), purified His-tagged AlkB protein released ethanol-soluble radioactive material from this methylated substrate (Fig. 1a). Similar activity was also observed with Flag-tagged AlkB protein (data not shown). The activity was dependent on α -ketoglutarate, inhibited by EDTA and stimulated by ascorbate (Fig. 1b). Inhibition by EDTA was overcome by adding an excess of Fe(NH₄)₂(SO₄)₂·6H₂O, thus demonstrating a requirement for Fe(II). At 100-fold higher AlkB concentrations, the effect of ascorbate was reduced. Thus, ascorbate was not essential but stimulated AlkB activity, probably by regenerating Fe(II) from Fe(III) (data not shown). The conditions for this assay were optimized to 50 mM HEPES-KOH, pH 8, 75 μ M Fe(NH₄)₂(SO₄)₂·6H₂O, 1 mM α -ketoglutarate, 2 mM ascorbate, and 50 μ g ml⁻¹ bovine serum albumin (BSA). To determine whether AlkB could also act on double-stranded DNA, the [¹⁴C]-labelled methylated poly(dA) was annealed to poly(dT). AlkB was approximately threefold more active on the double- compared with the single-stranded substrate (Fig. 1c).

To examine which methylated adenines were processed by AlkB, the poly(dA) substrate was incubated with AlkB, acid hydrolysed, and analysed by high-performance liquid chromatography (HPLC). In the poly(dA) polymer treated with [¹⁴C]methyl iodide, 1-methyladenine was the principal methylated base, 3-methyladenine was also abundant, but 7-methyladenine was a minor product¹⁴. After incubation with AlkB, the levels of 1-methyladenine

were reduced whereas the amounts of 3-methyladenine and 7-methyladenine remained unchanged (Fig. 2a). Thus, AlkB specifically catalysed removal of 1-methyladenine from methylated poly(dA). From the data in Fig. 1a, 0.1 pmol AlkB protein removed 1.7 pmol 1-methyladenine from methylated poly(dA) in 15 min, indicating that AlkB acts enzymatically and is not consumed during the reaction. This distinguishes its mode of action from O⁶-methylguanine-DNA methyltransferase.

A second modification that is formed to a greater extent in single-compared with double-stranded DNA is 3-methylcytosine. This lesion was also considered as a candidate substrate of AlkB. AlkB protein released radioactive material from poly(dC) treated with [¹⁴C]methyl iodide (data not shown). HPLC analysis showed that 3-methylcytosine was the only detectable modified base in this substrate¹⁴, and that it disappeared on incubation with AlkB protein (Fig. 2b). We have therefore identified two substrates of AlkB, 1-methyladenine and 3-methylcytosine, that are both generated in single-stranded DNA on treatment with S_N2 methylating agents.

The essential requirement for both α -ketoglutarate and Fe(II), inhibition by EDTA and stimulation by ascorbate strongly supports the proposal that AlkB is a dioxygenase that is dependent on α -ketoglutarate and Fe(II). Direct evidence that AlkB binds Fe(II) and α -ketoglutarate was obtained by examining the absorption spectrum of the anaerobic protein in the presence or absence of Fe(II) or α -ketoglutarate. The protein with bound metal and cofactor had an absorption peak at 500 nm (Fig. 1d). This chromophore is attributed to a weak charge transfer from Fe(II) to α -ketoglutarate, and is a spectroscopic signature of Fe(II)- and α -ketoglutarate-dependent dioxygenases^{16–18}. The AlkB absorption is slightly shifted compared with the 530-nm transition observed in other family members, and was not perturbed further on addition of methylated DNA.

We propose that AlkB repairs 1-methyladenine and 3-methylcytosine in DNA by oxidative demethylation. In such a mechanism, the lesions would be reverted to adenine and cytosine, formaldehyde would be generated and O₂ consumed. To demonstrate that AlkB directly reverts 1-methyladenine in DNA to adenine residues, a non-radioactive substrate was prepared in which 76% of the adenine residues were methylated to form 1-methyladenine. This was achieved by repeated treatments with DMS of an oligonucleotide containing adenine residues interspersed between inefficiently methylated thymine residues¹⁴. Only 4% of the adenines were recovered as 3-methyladenine, and 2% were 7-methyladenine (Fig. 2c). The low amount of 3-methyladenine might be caused by instability of the glycosyl bond and loss of this modification during

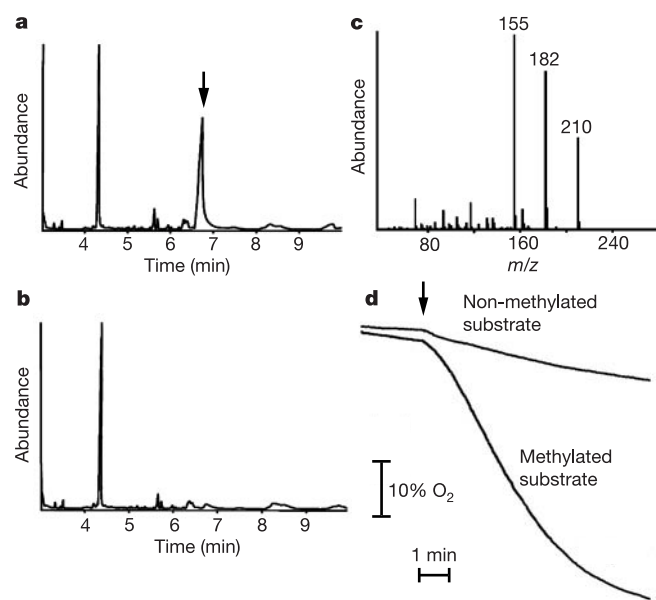


Figure 3 Production of formaldehyde and consumption of O₂ by AlkB. Methylated oligo(dA)•oligo(dT) was incubated with AlkB. The products were derivatized with PFPH and analysed by gas chromatography mass spectrometry. Gas chromatography traces are shown of derivatized products generated when the DNA was methylated (**a**) or not methylated (**b**). **c**, Mass spectrum of the product indicated by the arrow in **a**. **d**, Consumption of O₂ during incubation of AlkB with either methylated or non-methylated oligo(dA)•oligo(dT) was determined by using an oxygen electrode. AlkB (9 μ M) was added at the point indicated by the arrow.

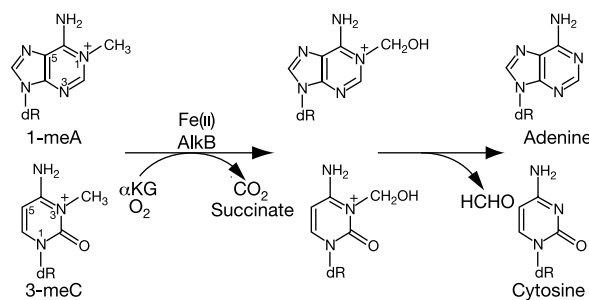


Figure 4 Schematic representation of the repair of 1-methyladenine (1-meA) and 3-methylcytosine (3-mec) residues in DNA by oxidative demethylation catalysed by AlkB. Oxidation of 1-methyladenine and 3-methylcytosine by AlkB requires O₂, α -ketoglutarate (α KG) and Fe(II), and generates CO₂ and succinate. Oxidized methyl groups are released as formaldehyde resulting in direct reversal of the lesions to the unmodified base residues. Note that the alkylation positions in 1-methyladenine and 3-methylcytosine are equivalent, although these sites in the pyrimidine rings of A and C are numbered differently by standard nomenclature. dR, deoxyribose.

the extensive DMS treatments. The heavily methylated substrate was incubated with AlkB in the optimized assay conditions, the DNA hydrolysed and individual bases quantified by HPLC and measurements of absorbance at 260 nm (A_{260}). In the presence of AlkB, a decrease in the amount of 1-methyladenine correlated with a stoichiometric increase in the amount of adenine recovered (Fig. 2c). We conclude that AlkB converts 1-methyladenine directly to adenine in DNA.

To determine whether formaldehyde (HCHO) was a reaction product, AlkB was incubated with MMS-treated poly(dA) annealed to poly(dT), the products were derivatized with pentafluorophenylhydrazine (PFPH), and analysed by gas chromatography. One derivatized product (arrow in Fig. 3a) arose only when the DNA substrate was methylated and Fe(II) and α -ketoglutarate were present (Fig. 3a, b), and had a mass spectrum identical to that for the HCHO-PFPH adduct¹⁹ (Fig. 3c). The release of HCHO was also monitored by a coupled spectrophotometric assay using formaldehyde dehydrogenase²⁰, and again was detected only in complete assay conditions with the heavily methylated DNA substrate (turn-over number 1.5 s^{-1} , data not shown). A yield of $140 \mu\text{M}$ HCHO correlated stoichiometrically with the amount of 1-methyladenine in the DNA substrate (7–10% of the adenines), the amount of O_2 consumed (Fig. 3d) and the succinate generated (data not shown). This stoichiometric relationship further verifies the proposal that AlkB is an α -ketoglutarate-dependent dioxygenase. In the absence of methylated DNA, an observed slow consumption of both oxygen (Fig. 3d) and α -ketoglutarate was consistent with a partial uncoupling of α -ketoglutarate decomposition and hydroxylation of the methylated DNA bases. Such uncoupling is a well known property of this family of enzymes^{21,22}.

Oxidative demethylation is an unprecedented mechanism of DNA repair. The proposed reaction mechanism by which AlkB repairs 1-methyladenine and 3-methylcytosine in DNA is shown in Fig. 4. Owing to the stability of the N–C bond in 1-methyladenine and 3-methylcytosine, demethylation by hydrolysis would be energetically unfavourable; consequently oxidative demethylation by reactive iron-oxygen species is required. Direct reversal of this base damage to unsubstituted parent residues would be a highly accurate form of DNA repair and agrees with *in vivo* observations that repair by AlkB is non-mutagenic⁵. *Escherichia coli* *alkB* mutants are more sensitive to alkylating agents during active growth than in stationary phase probably because 1-methyladenine and 3-methylcytosine are produced in single-stranded regions of DNA in replication forks and transcription bubbles⁵. DNA unfolds only transiently during replication and transcription, so it is beneficial that AlkB repairs its substrates not only in single strands but also, and even more efficiently, after DNA re-annealing (Fig. 1c). Of note, *Caulobacter crescentus* *alkB* expression is cell-cycle regulated with a pattern similar to activities required for DNA replication²³.

We have extended the family of α -ketoglutarate- and Fe(II)-dependent dioxygenases to include AlkB, as recently proposed on theoretical grounds⁶. These enzymes catalyse a variety of reactions including hydroxylations, desaturations and oxidative ring closures, and account for oxidation of proline in collagen, steps in the biosynthesis of several antibiotics and cellular metabolites, as well as biodegradation of selected compounds^{7,8}. No other members of this family are presently known to act on DNA, but it is notable that a fungal enzyme, thymine-7-hydroxylase, oxidizes the methyl group of free thymine²⁴. We suggest that oxidative demethylation requiring free-radical chemistry could be involved in other mechanisms of active removal of chemically stable adducts from DNA or histones, for example, the 5-methylcytosine demethylase activity^{25,26} that is apparently important in epigenetic control. □

Methods

AlkB protein assay and substrates

A total of 0.6 mg poly(dA) (average length 310 residues; Amersham Biosciences) was

treated with 1 mCi [^{14}C]methyl iodide (58 mCi mmol^{-1}) (Amersham Biosciences) in 0.8 ml^{-1} of 10 mM sodium cacodylate, pH 7, at 30°C for 6 h. The methylated polymer was precipitated in ethanol, dissolved in 10 mM Tris-HCl, pH 8, and had a specific activity of 1,860 counts per min (c.p.m.) per μg . To prepare a double-stranded substrate, [^{14}C]-labelled methylated poly(dA) was annealed to poly(dT) at 20°C . A total of 1.2 mg poly(dC) (average length 370 residues) in 1.3 ml of 50 mM HEPES-KOH, pH 8, was similarly treated with [^{14}C]methyl iodide to yield a specific activity of $700 \text{ c.p.m. } \mu\text{g}^{-1}$. His-tagged AlkB was overexpressed and purified as previously described by nickel-agarose affinity chromatography⁵, except that EDTA was added to the sonication buffer, and the extract was dialysed against the same buffer without EDTA. A plasmid coding for Flag-tagged AlkB was constructed by digestion of pBAR54 (ref. 5) with *Nde*I and *Nco*I to remove DNA coding for the His tag. This DNA fragment was replaced by a double-stranded oligonucleotide coding for the Flag tag. The Flag-tagged protein was purified by immunoaffinity chromatography (Sigma). Purified AlkB was incubated with the [^{14}C]methyl iodide-treated substrates in 50 mM HEPES-KOH, pH 8, $75 \mu\text{M}$ $\text{Fe}(\text{NH}_4)_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$, 1 mM α -ketoglutarate, 2 mM ascorbate, and $50 \mu\text{g ml}^{-1}$ BSA for 15 min at 37°C . The reaction was stopped by adding 11 mM EDTA, and the substrate was precipitated in ethanol in the presence of carrier calf thymus DNA. We monitored ethanol-soluble radioactive material by scintillation counting.

HPLC analysis

Methylated adenine residues were released from [^{14}C]-labelled methylated poly(dA) by hydrolysis in 0.1 M HCl at 95°C for 1 h; 3-methylcytosine was released from methylated poly(dC) by treatment with 90% formic acid at 180°C for 20 min. Methylated bases were analysed by HPLC on a Whatman Partisil 10 cation exchange column in 0.1 M ammonium formate, pH 3.6. A gradient of MeOH from 20 to 40% was applied to separate the methylated adenine derivatives, and a gradient from 5 to 40% MeOH was used to analyse 3-methylcytosine.

High-level methylation of an oligonucleotide

A 41-base oligonucleotide (TTTTT(ATTTTT)₅) was treated with 50 mM DMS in 75 mM sodium cacodylate, pH 7.4, at 30°C , twice for 2 h, and then four times for 1 h. Between each treatment the DMS was removed by centrifugation through a G25 Sephadex mini-column equilibrated in the same buffer. The level of methylation of adenine residues was examined by acid hydrolysis, HPLC and A_{260} measurements. The A_{260} ratio of adenine to 1-methyladenine was 1.04, which was determined by monitoring known amounts of these purines in the same conditions.

Gas chromatography mass spectrometry

Oligo(dA) (25 residues) was treated with 500 mM MMS at 30°C for 30 min and, after removal of the MMS, annealed to oligo(dT). AlkB ($38 \mu\text{M}$) was incubated with this substrate ($50 \mu\text{M}$) in 50 mM HEPES-KOH, pH 8, containing 1 mM α -ketoglutarate, $75 \mu\text{M}$ $\text{Fe}(\text{NH}_4)_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$, and 2 mM ascorbate in a final volume of 200 μl . After 10 min at 30°C , 200 μl 15 mM PFPH in 1.2 M phosphoric acid was added, and incubation continued at 50°C for 2 h. The reaction mixtures were extracted with 100 μl of 85:15 hexane:dichloromethane, and 20 μl was analysed by gas chromatography mass spectrometry as described¹⁹.

O_2 consumption

We mixed 50 mM HEPES, pH 8, with 1 mM α -ketoglutarate, $75 \mu\text{M}$ $\text{Fe}(\text{NH}_4)_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$, 100 μM ascorbate, and $50 \mu\text{M}$ oligonucleotide substrate, prepared as described for gas chromatography mass spectrometry, for 3 min in a YSI 5300 oxygen electrode at 30°C to allow equilibration with atmospheric oxygen. The electrode was precalibrated with air-saturated water ($236 \mu\text{M O}_2$). AlkB was added through a gas-tight syringe to a final concentration of $9 \mu\text{M}$.

Received 7 March; accepted 6 June 2002; doi:10.1038/nature00908.

- Lindahl, T., Sedgwick, B., Sekiguchi, M. & Nakabeppu, Y. Regulation and expression of the adaptive response to alkylating agents. *Annu. Rev. Biochem.* **57**, 133–157 (1988).
- Seeberg, E. & Berdal, K. G. In *Base Excision Repair of DNA Damage: Repair of Alkylation Damage to DNA* (ed. Hickson, I. D.) 151–168 (Landes Bioscience, Austin, 1999).
- Kataoka, H., Yamamoto, Y. & Sekiguchi, M. A new gene (*alkB*) of *Escherichia coli* that controls sensitivity to methyl methane sulfonate. *J. Bacteriol.* **153**, 1301–1307 (1983).
- Kondo, H. et al. Structure and expression of the *alkB* gene of *Escherichia coli* related to the repair of alkylated DNA. *J. Biol. Chem.* **261**, 15772–15777 (1986).
- Dinglay, S., Treweek, S. C., Lindahl, T. & Sedgwick, B. Defective processing of methylated single-stranded DNA by *E. coli* *alkB* mutants. *Genes Dev.* **14**, 2097–2105 (2000).
- Aravind, L. & Koonin, E. V. The DNA-repair protein AlkB, Egl-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. *Genome Biol.* **2**, RESEARCH0007 (2001).
- Prescott, A. G. & Lloyd, M. D. The iron(II) and 2-oxoacid-dependent dioxygenases and their role in metabolism. *Nat. Prod. Rep.* **17**, 367–383 (2000).
- Ryle, M. J. & Hausinger, R. P. Non-heme iron oxygenases. *Curr. Opin. Chem. Biol.* **6**, 193–201 (2002).
- Rebeck, G. W. & Samson, L. Increased spontaneous mutation and alkylation sensitivity of *Escherichia coli* strains lacking the *ogt* O⁶-methylguanine-DNA methyltransferase. *J. Bacteriol.* **173**, 2068–2076 (1991).
- Vaughan, P., Sedgwick, B., Hall, J., Gannon, J. & Lindahl, T. Environmental mutagens that induce the adaptive response to alkylating agents in *Escherichia coli*. *Carcinogenesis* **12**, 263–268 (1991).
- Chen, B. J., Carroll, P. & Samson, L. The *Escherichia coli* AlkB protein protects human cells against alkylation-induced toxicity. *J. Bacteriol.* **176**, 6255–6261 (1994).
- Wei, Y., Carter, K. C., Wang, R. & Shell, B. K. Molecular cloning and functional analysis of a human cDNA encoding an *Escherichia coli* AlkB homolog, a protein involved in DNA alkylation damage repair. *Nucleic Acids Res.* **24**, 931–937 (1996).

13. Bodell, W. J. & Singer, B. Influence of hydrogen bonding in DNA and polynucleotides on reaction of nitrogens and oxygens toward ethylnitrosourea. *Biochemistry* **18**, 2860–2863 (1979).
14. Singer, B. & Grunberger, D. *Molecular Biology of Mutagens and Carcinogens: Reactions of Directly Acting Agents with Nucleic Acids* 45–96 (Plenum, New York, 1983).
15. Aravind, L., Walker, D. R. & Koonin, E. V. Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.* **27**, 1223–1242 (1999).
16. Hegg, E. L. *et al.* Herbicide-degrading α -keto acid-dependent enzyme: metal coordination environment and mechanistic insights. *Biochemistry* **38**, 16714–16726 (1999).
17. Pavel, E. G. *et al.* Circular dichroism and magnetic circular dichroism spectroscopic studies of the non-heme ferrous active site in clavaminase synthase and its interaction with α -ketoglutarate co-substrate. *J. Am. Chem. Soc.* **120**, 743–753 (1998).
18. Ryle, M. J., Padmakumar, R. & Hausinger, R. P. Stopped-flow kinetic analysis of *Escherichia coli* taurine/ α -ketoglutarate dioxygenase: interactions with α -ketoglutarate, taurine, and oxygen. *Biochemistry* **38**, 15278–15286 (1999).
19. Heck, H. D. A., White, E. L. & Cassanova-Schmitz, M. Determination of formaldehyde in biological tissues by gas chromatography/mass spectrometry. *Biomed. Mass Spectrom.* **9**, 347–353 (1982).
20. Lizcano, M. J., Unzeta, M. & Tipton, K. F. A spectrophotometric method for determining the oxidative deamination of methylamine by amine oxidases. *Anal. Biochem.* **286**, 75–79 (2000).
21. de Jong, L., Albracht, S. P. & Kemp, A. Prolyl 4-hydroxylase activity in relation to the oxidation state of enzyme-bound iron. The role of ascorbate in peptidyl proline hydroxylation. *Biochim. Biophys. Acta* **704**, 326–332 (1982).
22. Myllyla, R., Majamaa, K., Gunzler, V., Hanauske-Abel, H. M. & Kivirikko, K. I. Ascorbate is consumed stoichiometrically in the uncoupled reactions catalysed by prolyl 4-hydroxylase and lysyl hydroxylase. *J. Biol. Chem.* **259**, 5403–5405 (1984).
23. Colombi, D. & Gomes, S. L. An *alkB* homologue is differentially transcribed during the *Caulobacter crescentus* cell cycle. *J. Bacteriol.* **179**, 3139–3145 (1997).
24. Thornburg, L. D., Lai, M.-T., Wishnok, J. S. & Stubbe, J. A non-heme iron protein with heme tendencies: an investigation of the substrate specificity of thymine hydroxylase. *Biochemistry* **32**, 14023–14033 (1993).
25. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501–502 (2000).
26. Smith, S. S. Gilbert's conjecture: the search for DNA (cytosine-5) demethylases and the emergence of new functions for eukaryotic DNA (cytosine-5) methyltransferases. *J. Mol. Biol.* **302**, 1–7 (2000).

Acknowledgements

We thank J. McCracken for advice, and T. Duncan and P. Robins for technical help. This work was supported by Cancer Research UK, and by NIH grants to R.P.H. and J. McCracken.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (e-mail: b.sedgwick@cancer.org.uk) or R.P.H. (e-mail: hausinge@msu.edu).

AlkB-mediated oxidative demethylation reverses DNA damage in *Escherichia coli*

Pål Ø. Farnes, Rune F. Johansen & Erling Seeberg

Centre for Molecular Biology and Neuroscience, and Institute of Medical Microbiology, University of Oslo, National Hospital, 0027 Oslo, Norway

The bacterial AlkB protein is known to be involved in cellular recovery from alkylation damage; however, the function of this protein remains unknown. AlkB homologues have been identified in several organisms, including humans, and a recent sequence alignment study has suggested that these proteins may belong to a superfamily of 2-oxoglutarate-dependent and iron-dependent oxygenases (2OG-Fe(II)-oxygenases)¹. Here we show that AlkB from *Escherichia coli* is indeed a 2-oxoglutarate-dependent and iron-dependent DNA repair enzyme that releases replication blocks in alkylated DNA by a mechanism involving oxidative demethylation of 1-methyladenine residues. This mechanism represents a new pathway for DNA repair and the third type of DNA damage reversal mechanism so far discovered.

Alkylating agents are abundant in nature and introduce serious damage to the DNA of living organisms. To prevent the lethal and

mutagenic effects of such damage, repair mechanisms have evolved that are highly conserved throughout evolution. The mechanisms of alkylation repair were first elucidated in *E. coli* and have been shown to involve DNA glycosylases for removal of cytotoxic N-alkylated purines (for example, 3-methyladenine) and alkyltransferases for removal of premutagenic O-alkylated bases (for example, O⁶-methylguanine). In the transferase reaction the alkyl group is transferred from the DNA to the transferase itself, which is then consumed during repair in a stoichiometric fashion (reviewed in ref. 2). The elucidation of these processes was aided greatly by the isolation of alkylation-sensitive mutants, which led to the identification of *tag* and *alkA* as genes encoding DNA glycosylases^{3,4}, and *ada* and *ogt* coding for alkyltransferases⁵. Another alkylation repair gene, *alkB*, was also identified early during this work⁶. However, the precise function of the AlkB protein has yet to be resolved despite the fact that AlkB homologues have been identified in a number of organisms, including *Schizosaccharomyces pombe*, *Drosophila*, *Caenorhabditis elegans* and humans^{1,7}, suggesting an important and universal function for this gene product.

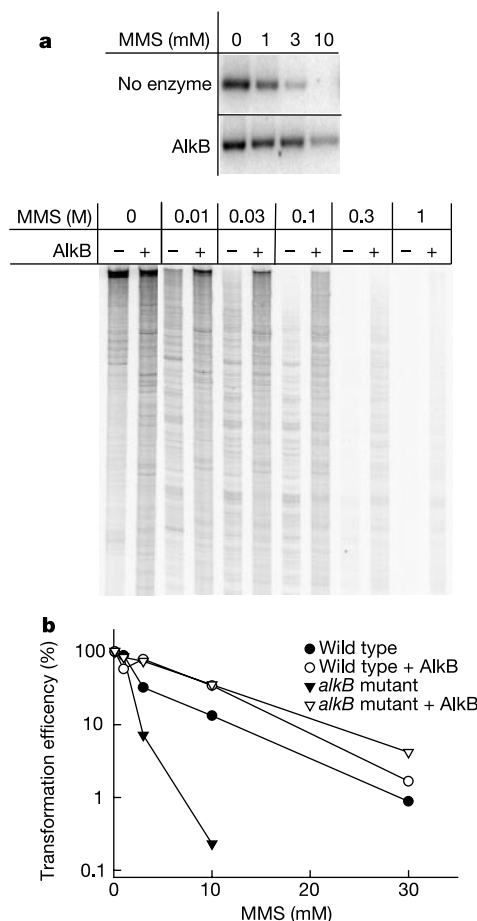


Figure 1 Repair of MMS-treated M13mp18 ssDNA by treatment with AlkB. **a**, Release of replication blocks in M13 ssDNA. Top panel: M13 ssDNA was treated with the indicated concentrations of MMS, and used in a replication assay with *E. coli* protein extracts in the presence of [α -³²P]dCTP, 2-oxoglutarate, FeSO₄ and AlkB protein (when indicated). The replication products were analysed by agarose gel electrophoresis, and the relaxed form (RFII) of M13 dsDNA is shown. Bottom panel: MMS-treated DNA was incubated with 2-oxoglutarate, FeSO₄ and AlkB protein (when indicated), and used in a primer extension assay with T7 DNA polymerase in the presence of [α -³⁵S]dATP. The extension products were analysed by denaturing polyacrylamide gel electrophoresis. **b**, Transformation efficiency of *E. coli* wild type (circles) and *alkB* mutant (triangles) by MMS-treated M13 ssDNA incubated with (open symbols) or without (filled symbols) AlkB.

Recently, two studies have provided important clues to the function of AlkB. First, ref. 8 used a host cell reactivation assay with methyl methanesulphonate (MMS)-treated filamentous phage M13 to show that AlkB is particularly important for the rescue of alkylated single-stranded DNA (ssDNA). Second, iterative searches in protein databases have suggested that AlkB is a member of the superfamily of 2-oxoglutarate- and iron-dependent oxygenases (2OG-Fe(II)-oxygenases)¹. In this study, we have elaborated on these findings and used a system for M13 DNA replication in cell extracts to show that AlkB releases replication blocks from MMS-treated ssDNA in reaction mixtures containing the cofactors Fe²⁺ and 2-oxoglutarate (Fig. 1a, upper panel). Furthermore, the transformation efficiency of an *alkB* mutant by methylated M13 ssDNA was greatly enhanced by treatment of the DNA with purified AlkB in the presence of Fe²⁺ and 2-oxoglutarate (Fig. 1b), implying that AlkB is able to exert its function in the absence of other proteins. AlkB pretreatment was also found to release replication stops in MMS-treated M13 ssDNA in a primer extension reaction with T7 DNA polymerase (Fig. 1a, lower panel). When the extension products were analysed by denaturing gel electrophoresis along with DNA sequencing reactions, the pattern of stop sites released by AlkB corresponded, at least partly, to adenine lesions (data not shown).

To characterize further the nature of the lesions repaired by AlkB, single-stranded poly(dA) was methylated with *N*-[³H]methyl-*N*-nitrosourea (MNU) and incubated with AlkB and the cofactors. A

substantial fraction of the radioactivity incorporated was released in an ethanol soluble form from the poly(dA) ssDNA by AlkB (Fig. 2a; see also Supplementary Information Fig. I). No significant release of radioactivity was observed from double-stranded homopolymeric poly(dA)•poly(dT) alkylated under similar conditions. When [³H]MNU-treated M13 ssDNA was used, about 5% of the total radioactivity was released by AlkB treatment (Fig. 2a). AlkB was also capable of releasing radioactivity from double-stranded DNA (dsDNA) that was generated by the methylation of an A-rich oligonucleotide in the single-stranded form, followed by annealing to the complementary strand (Fig. 2a). However, compared with the single-stranded substrate, approximately tenfold more AlkB protein was required to reach the same level of excision (Fig. 2a). Therefore, although the lesions targeted by AlkB are introduced primarily into ssDNA, AlkB is also capable of removing lesions from dsDNA.

The AlkB-mediated release of radioactivity from [³H]methyl poly(dA) was measured in the presence of varying concentrations of Fe²⁺ and 2-oxoglutarate, and the results showed that approximately 3 μM Fe²⁺ and 10 μM 2-oxoglutarate were required to obtain full activity of AlkB (apparent Michaelis constant (*K_m*) values of 0.2 and 1 μM, respectively; Fig. 2b, c). Many 2OG-Fe(II)-oxygenases have a requirement for ascorbate, and we also tested the effect of ascorbate on the activity of AlkB. We found that

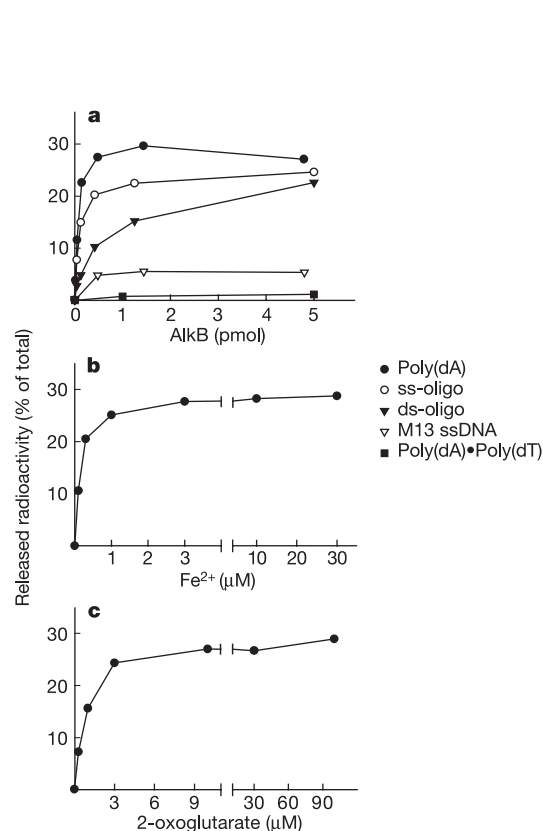


Figure 2 AlkB-mediated release of radioactivity from [³H]methyl DNA. **a**, AlkB activity on poly(dA) (filled circles), a single-stranded, A-rich oligonucleotide (AAAGCAAGAACGAAAAAGCGAAA) (open circles), the same oligonucleotide associated with its (unmethylated) complementary strand (filled triangles), M13 ssDNA (open triangles) and poly(dA)•poly(dT) (filled squares), in the presence of 100 μM 2-oxoglutarate and 40 μM Fe²⁺. **b**, Fe²⁺ dependence of AlkB activity. The reactions contained 200 ng methylated poly(dA), 1 pmol AlkB, and 100 μM 2-oxoglutarate in standard mixtures. **c**, 2-oxoglutarate dependence of AlkB activity. The reactions contained 200 ng methylated poly(dA), 1 pmol AlkB, and 30 μM Fe²⁺ in standard mixtures.

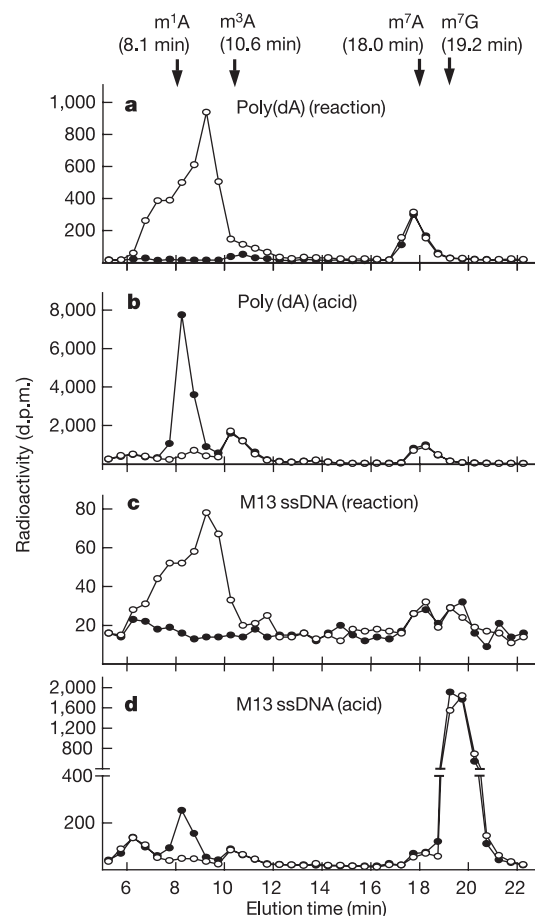


Figure 3 Reverse phase HPLC elution profiles of radioactivity released from [³H]methyl ssDNA by AlkB and/or acid treatment. **a**, **c**, HPLC profiles of material recovered in an ethanol soluble form from [³H]methyl DNA as indicated after incubations with (open symbols) or without (filled symbols) AlkB. **b**, **d**, HPLC profiles of methylpurines released by acid treatment of ethanol precipitated DNA after incubations with (open symbols) or without (filled symbols) AlkB. Arrows indicate the elution time of 1-methyladenine (m¹A), 3-methyladenine (m³A), 7-methyladenine (m⁷A) and 7-methylguanine (m⁷G).

ascorbate only caused a marginal stimulation of the activity of AlkB (Supplementary Information Fig. II).

When [^3H]methyl dsDNA is incubated with 3-methyladenine DNA glycosylases and the ethanol soluble radioactivity in the supernatant is analysed, methylated bases are identified. However, when the supernatant after AlkB treatment of [^3H]methyl poly(dA) was subjected to reverse phase high-performance liquid chromatography (HPLC), a broad peak was detected that did not correspond to 1-methyladenine, 3-methyladenine or 7-methyladenine—the principal lesions formed in poly(dA) by chemical methylation (Fig. 3a). To quantify any changes in the amount of alkylated base residues present in the ethanol-precipitated DNA before and after AlkB treatment, the precipitated DNA was treated with HCl acid to

release purines, and the released material was then analysed by HPLC. The results showed that the peak of radioactivity corresponding to 1-methyladenine disappeared after the AlkB treatment, whereas both 3-methyladenine and 7-methyladenine were unaffected (Fig. 3b). It thus appears that AlkB mediates the release of the radioactivity contained in 1-methyladenine residues, but that this release does not involve excision of the intact base. Similar results were obtained when methylated M13 ssDNA was used as substrate (Fig. 3c, d).

To characterize further the products of the AlkB reaction, poly(dA) was subjected to three consecutive treatments with an elevated concentration (300 mM) of non-radioactive dimethylsulphate (DMS) to produce a high content of 1-methyladenine in the DNA. The amount of 1-methyladenine remaining in the DNA could then be quantified by ultraviolet absorption of the HPLC eluate (Fig. 4a). When this heavily methylated DNA was subjected to AlkB treatment followed by ethanol precipitation, acid hydrolysis, and HPLC of the purines released, the 1-methyladenine peak was strongly reduced, whereas the adenine peak increased concomitantly from 62% to 87% of the total (Fig. 4a, compare top and bottom panels). The correspondence between the decrease in 1-methyladenine and concomitant increase in adenine was also found to be valid for other doses of DMS exposure (data not shown). These data show that AlkB is a damage reversal enzyme and restores adenine from 1-methyladenine by demethylation.

OG-Fe(II)-oxygenases share some mechanistic features with the cytochrome P450 enzymes, which are involved in the detoxification of foreign substances (reviewed in ref. 9). Both groups of enzymes require ferrous iron and catalyse hydroxylation reactions where O_2 is the oxygen donor (Fig. 4c, d). Interestingly, when *N*-methyl groups ($\text{R}_1\text{R}_2\text{N}-\text{CH}_3$) are hydroxylated by cytochrome P450, the resulting $\text{R}_1\text{R}_2\text{N}-\text{CH}_2\text{OH}$ is spontaneously converted to $\text{R}_1\text{R}_2\text{N}-\text{H}$ and formaldehyde (HCHO), with demethylation being the net result (Fig. 4c). To investigate whether AlkB would act by a similar mechanism, we measured whether formaldehyde was generated in the reaction mixture during incubation of methyl-poly(dA) ssDNA with AlkB. Using the Nash assay^{10,11} to measure formaldehyde, we found that formaldehyde was formed during the reaction, and that the quantity of formaldehyde formed corresponded stoichiometrically to the amount of 1-methyladenine demethylated in a parallel reaction (Fig. 4b). Using the same assay, we found that formaldehyde produced a broad peak after HPLC, eluting at a position corresponding to the broad peak of radioactivity recovered after HPLC of the ethanol supernatants following AlkB treatment of [^3H]methyl DNA (data not shown). From the amounts of formaldehyde released and 1-methyladenine demethylated, we calculated that one molecule of AlkB mediated the release of 6.0 methyl groups in 30 min, which corresponds to an enzyme turnover of 0.2 min^{-1} , indicating that AlkB is a true enzyme. However, it should be noted that the turnover rate for the radiolabelled substrate (Fig. 2a) is much lower, probably because of an inhibitory effect of the large excess of normal bases in this substrate, which only contains one 1-methyladenine residue for every 30,000 bases. As predicted from the reaction mechanism for the 2OG-Fe(II)-oxygenases, succinate was formed during the AlkB reaction (Supplementary Information Fig. III). Significant amounts of succinate were generated also in the absence of DNA ('uncoupled reaction'), but the addition of substrate caused an increase in succinate formation corresponding to the amount of 1-methyladenine in the substrate. Altogether, our data provide evidence that the AlkB protein is an oxidative demethylase acting on 1-methyladenine residues in DNA. The demethylation probably occurs in two steps; that is, by an AlkB catalysed hydroxylation of the methyl group, followed by a spontaneous release of formaldehyde (Fig. 4d).

Originally, the *alkB* mutant was reported to be sensitive to $\text{S}_{\text{N}}2$ alkylating agents such as MMS and DMS but not to $\text{S}_{\text{N}}1$ agents such as *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine (MNNG) and MNU^{6,12}.

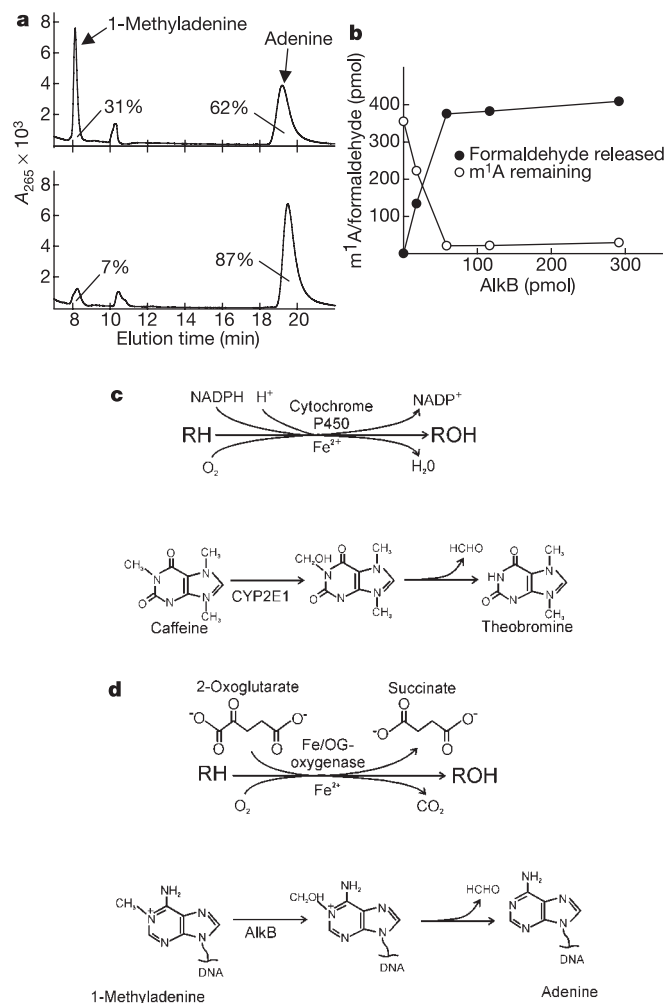


Figure 4 Oxidative demethylation of 1-methyladenine lesions in DNA by the AlkB protein. **a**, Conversion of 1-methyladenine to adenine by AlkB. Methylated poly(dA) was incubated in the absence (top) or presence (bottom) of AlkB, treated with acid to release purine bases, and the released material was analysed by HPLC. A_{265} , absorbance at 265 nm. **b**, Quantification of formaldehyde release (filled circles) and reduction in 1-methyladenine content (open circles) after the incubation of AlkB protein with methylated poly(dA) (2.5 nmol of nucleotides, containing 14% 1-methyladenine). Parallel experiments were performed where formaldehyde was measured by the Nash assay, and the 1-methyladenine content was determined by HPLC analysis of the acid hydrolysate of the DNA. **c**, Basic mechanism for hydroxylation by cytochrome P450 (CYP) enzymes (top) and example of a demethylation reaction catalysed by a CYP enzyme (bottom). **d**, Basic mechanism for hydroxylation by 2OG-Fe(II)-oxygenases (top) and proposed mechanism for demethylation of 1-methyladenine by AlkB (bottom).

This can be ascribed to a much lower amount of 1-methyladenine induced in DNA by MNNG and MNU than by MMS and DMS¹³. The finding that AlkB is much more important for the recovery of ssDNA⁸ can be explained by the much higher efficiency with which 1-methyladenine is produced in single-stranded versus double-stranded DNA by all these agents¹³. [³H]MNU rather than radio-active MMS or DMS was used in our experiments because of availability; however, significant amounts of 1-methyladenine are also produced by MNU in ssDNA (Fig. 3b, d). The fact that ssDNA is much more vulnerable to the induction of 1-methyladenine by chemical methylation than dsDNA suggests that AlkB, *in vivo*, may be located together with the replication machinery at the replication fork or at sites of transcription; that is, in regions that are transiently single-stranded. Nevertheless, a certain amount of 1-methyladenine will also be formed by methylation of dsDNA, and AlkB may also be required for global genome repair.

Oxidative DNA demethylation represents a third mechanism for repairing alkylation damage to DNA, which combines favourable features of the other two mechanisms described so far. Similar to the alkylguanine DNA alkyltransferases, AlkB repairs lesions without the participation of other protein factors, and, as with the 3-methyladenine DNA glycosylases, AlkB is not consumed during the reaction and is, therefore, a true enzyme. Moreover, the AlkB-mediated reaction represents the third type of DNA damage reversal mechanism so far described; different from the demethylation performed by the DNA alkyltransferases and the ultraviolet DNA photolyase cleavage of pyrimidine dimers. A number of different AlkB homologues seem to exist in human cells and other organisms¹, and one might speculate that the AlkB family of proteins could perform several different dealkylation as well as other oxidation reactions on modified nucleic acids. □

Methods

Strains

The F⁺-episome was introduced into *E. coli* strains AB1157 (wild type) and HK82 (*alkB*) as described⁸ to generate the strains AB1157/F⁺ and HK82/F⁺, respectively.

Expression and purification of the AlkB protein

We expressed and purified AlkB protein by standard techniques¹⁴. In brief, the *alkB* gene was amplified by polymerase chain reaction (PCR) from DNA isolated from the *E. coli* strain W3110 using primers that allowed cloning of *alkB* between the *Bam*HI and *Eco*RI sites of the plasmid pGEX-2T (Amersham-Pharmacia Biotech). The resulting plasmid was transformed into the *E. coli* strain BL21 Star(DE3)pLysS (Invitrogen), and a fusion protein between AlkB and glutathione S-transferase was expressed (GST-AlkB). The GST-AlkB fusion protein was purified from *E. coli* protein extracts by affinity chromatography on a GSTrap (Amersham Pharmacia Biotech) column. In the GST-AlkB fusion protein, GST and AlkB are separated by a thrombin cleavage site; the AlkB protein was obtained by on-column cleavage of GST-AlkB with thrombin.

Preparation of methylated DNA

N-[³H]methyl-N-nitrosourea (specific activity 718 GBq mmol⁻¹) treated DNA was prepared as described previously¹⁵. The radioactivity incorporated into the substrates was typically 20,000 disintegrations per minute (d.p.m.) per µg, and 0.1 µg DNA was typically used per reaction when AlkB-mediated release of radioactivity was studied. DNA was treated with DMS or MMS in 100 mM sodium cacodylate, pH 7.0, for 30 min at 30 °C, followed by ethanol precipitation of the DNA, and resuspension in 10 mM Tris, pH 8.

M13 DNA *in vitro* replication by *E. coli* extracts

Extract preparation and the replication reactions were performed essentially as described¹⁶, except that ascorbate (2 mM), 2-oxoglutarate (1 mM), FeSO₄ (0.1 mM) and in some cases AlkB (10 pmol) was added to the replication reactions. Replication products were separated on agarose gels, which were subsequently dried and subjected to phosphorimaging.

Repair of methylated DNA by AlkB protein

The reaction mixture (50–100 µl) typically contained methylated DNA (200 ng), AlkB (5 pmol), 50 mM KCl, 50 mM Tris (pH 8), 10 mM Mg(OAc)₂, 1 mM dithiothreitol, 100 µg ml⁻¹ bovine serum albumin (BSA), 100 µM 2-oxoglutarate, and 40 µM FeSO₄, and was incubated for 30 min at 37 °C. In some experiments, ascorbate (2 mM) was present, but was later found to be unnecessary and subsequently omitted.

Primer extension by T7 DNA polymerase

The reactions were identical to standard Sequenase (Amersham Pharmacia Biotech) DNA sequencing reactions (Sequenase Version 2.0 DNA Sequencing Kit;

<http://www.usbweb.com>) except that dideoxynucleotides were omitted from the extension mixture. A standard M13 sequencing primer (0.25 pmol) was annealed to MMS-treated M13 ssDNA (200 ng) that had been incubated under standard repair conditions in the absence or presence of AlkB protein. The primed template was first incubated with Sequenase (0.8 U) for 5 min at room temperature in the presence of limiting amounts (0.1 µM) of [α-³⁵S]dATP, dCTP, dGTP and dTTP, to provide labelling of the extension products. Subsequently, dNTPs (10 µM of each) were added, and the reaction was incubated for 15 min at 37 °C, allowing further extension. The extension products were separated on a 15% denaturing polyacrylamide gel, and detected by phosphorimaging of the dried gel.

Transformation of *E. coli* by MMS-treated M13mp18 ssDNA

Competent *E. coli* cells were prepared by using a modification¹⁷ of the one-step method described previously¹⁸. Incubation of MMS-treated M13 ssDNA (500 ng) in the absence or presence of AlkB (2 pmol) was performed in a 20-µl reaction under standard conditions, except that the concentration of Fe₂SO₄ was 10 µM. After the reaction the ssDNA was recovered by ethanol precipitation, and 200 ng was used for each transformation.

Analysis of products released from methylated DNA by AlkB

Repair reactions were terminated by ethanol precipitation of the DNA, and the resulting supernatant was either analysed directly by scintillation counting, or subjected to reverse-phase HPLC under conditions described previously¹⁹, using two Chrompack Inertsil 5 ODS-2 150 × 4 mm columns in series, in some cases followed by scintillation counting of the fractions. The DNA pellet resulting from ethanol precipitation of the repair reaction was in some cases subjected to acid treatment (0.1 M HCl for 20 min at 70 °C) to release purine bases. The depurinated DNA was removed by ethanol precipitation and the released material subjected to reverse-phase HPLC. Formaldehyde release was measured by using a fluorimetric assay¹¹ based on the reaction between the Nash reagent and formaldehyde¹⁰, generating diacetyldihydrolutidine, which was detected by using an excitation wavelength of 405 nm and measuring emission at 510 nm. Briefly, poly(dA) (approximately 1 µg) that had been heavily methylated through repeated DMS treatment was subjected to AlkB repair under slightly modified conditions; that is, Tris was replaced with HEPES, the concentration of dithiothreitol was 100 µM, the concentration of Fe₂SO₄ was 20 µM and BSA was omitted. The repair reaction was diluted directly with the Nash reagent, and formaldehyde determined as described¹¹.

Received 27 May; accepted 30 July 2002; doi:10.1038/nature01048.

- Aravind, L. & Koonin, E. V. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. *Genome Biol.* **2**, 0007.1–0007.8 (2001).
- Pegg, A. E. Repair of O(6)-alkylguanine by alkyltransferases. *Mutat. Res.* **462**, 83–100 (2000).
- Evensen, G. & Seeberg, E. Adaptation to alkylation resistance involves the induction of a DNA glycosylase. *Nature* **296**, 773–775 (1982).
- Karran, P., Hjelmgren, T. & Lindahl, T. Induction of a DNA glycosylase for N-methylated purines is part of the adaptive response to alkylating agents. *Nature* **296**, 770–773 (1982).
- Karran, P., Lindahl, T. & Griffin, B. Adaptive response to alkylating agents involves alternation *in situ* of O6-methylguanine residues in DNA. *Nature* **280**, 76–77 (1979).
- Kataoka, H., Yamamoto, Y. & Sekiguchi, M. A new gene (*alkB*) of *Escherichia coli* that controls sensitivity to methyl methane sulfonate. *J. Bacteriol.* **153**, 1301–1307 (1983).
- Wei, Y. F., Carter, K. C., Wang, R. P. & Shell, B. K. Molecular cloning and functional analysis of a human cDNA encoding an *Escherichia coli* AlkB homolog, a protein involved in DNA alkylation damage repair. *Nucleic Acids Res.* **24**, 931–937 (1996).
- Dinglay, S., Treweek, S. C., Lindahl, T. & Sedgwick, B. Defective processing of methylated single-stranded DNA by *E. coli* AlkB mutants. *Genes Dev.* **14**, 2097–2105 (2000).
- Parkinson, A. *Casarett and Doull's Toxicology* (ed. Claassen, C. D.) 113–186 (McGraw-Hill, New York, 1996).
- Nash, T. The colorimetric estimation of formaldehyde by means of the Hantzsch reaction. *Biochem. J.* **55**, 416–421 (1953).
- Rapaport, R., Hanukoglu, I. & Sklan, D. A fluorimetric assay for hydrogen peroxide, suitable for NAD(P)H-dependent superoxide generating redox systems. *Anal. Biochem.* **218**, 309–313 (1994).
- Chen, B. J., Carroll, P. & Samson, L. The *Escherichia coli* AlkB protein protects human cells against alkylation-induced toxicity. *J. Bacteriol.* **176**, 6255–6261 (1994).
- Singer, B. & Grunberger, D. *Molecular Biology of Mutagens and Carcinogens* (Plenum, New York, 1983).
- Amersham Pharmacia Biotech *GST Gene Fusion System: User Manual*; (1997) (www.amershambiosciences.com).
- Bjelland, S., Björås, M. & Seeberg, E. Excision of 3-methylguanine from alkylated DNA by 3-methyladenine DNA glycosylase I of *Escherichia coli*. *Nucleic Acids Res.* **21**, 2045–2049 (1993).
- Wang, G., Rahman, M. S. & Humayun, M. Z. Replication of M13 single-stranded viral DNA bearing single site-specific adducts by *Escherichia coli* cell extracts: differential efficiency of translation DNA synthesis for SOS-dependent and SOS-independent lesions. *Biochemistry* **36**, 9486–9492 (1997).
- Palejwala, V. A., Sinha, D. & Humayun, M. Z. Mechanisms of mutagenesis by exocyclic DNA adducts. Transfection of M13 viral DNA bearing a site-specific adduct shows that ethenocytosine is a highly efficient RecA-independent mutagenic noninstructive lesion. *Biochemistry* **30**, 8736–8743 (1991).
- Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution. *Proc. Natl Acad. Sci. USA* **86**, 2172–2175 (1989).
- Hofer, T. & Moller, L. Reduction of oxidation during the preparation of DNA and analysis of 8-hydroxy-2'-deoxyguanosine. *Chem. Res. Toxicol.* **11**, 882–887 (1998).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements

We thank M. Björås, L. Eide, K. Baynton, K. I. Kristiansen, J. Myllyharju, K. Skarstad and J. Klavness for help and discussions, and L. Eide, K. Baynton and A. Klungland for critical reading of the manuscript. We are grateful to M. Björås for preparation of [³H]MNU-labelled DNA substrates, to L. Eide for the construction of the plasmid for expression of GST-AlkB, and to D. Daoudi for technical assistance. This work was supported by the Research Council of Norway and the Norwegian Cancer Society. E.S. also acknowledges support from the European Commission.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.Ø.F. (e-mail: pal.falnes@labmed.uio.no).

Comprehensive proteomic analysis of the human spliceosome

Zhaolan Zhou^{*†‡}, Lawrence J. Licklider^{†‡}, Steven P. Gygi^{†‡} & Robin Reed[†]

^{*} Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

[†] Department of Cell Biology, [‡] Taplin Biological Mass Spectrometry Facility, Harvard Medical School, Boston, Massachusetts 02115, USA

The precise excision of introns from pre-messenger RNA is performed by the spliceosome, a macromolecular machine containing five small nuclear RNAs and numerous proteins. Much has been learned about the protein components of the spliceosome from analysis of individual purified small nuclear ribonucleoproteins¹ and salt-stable spliceosome 'core' particles^{2,3}. However, the complete set of proteins that constitutes intact functional spliceosomes has yet to be identified. Here we use maltose-binding protein affinity chromatography^{4,5} to isolate spliceosomes in highly purified and functional form. Using nanoscale microcapillary liquid chromatography tandem mass spectrometry⁶, we identify ~145 distinct spliceosomal proteins, making the spliceosome the most complex cellular machine so far characterized. Our spliceosomes comprise all previously known splicing factors and 58 newly identified components. The spliceosome contains at least 30 proteins with known or putative roles in gene expression steps other than splicing. This complexity may be required not only for splicing multi-intronic metazoan pre-messenger RNAs, but also for mediating the extensive coupling between splicing and other steps in gene expression.

Spliceosomes undergo multiple assembly stages and conformational changes during the splicing reaction^{7–9}. To obtain as complete a spliceosomal proteome as possible, we purified spliceosomes from a mixture of all stages of assembly, including spliceosomes that had undergone the first or second catalytic steps of the splicing reaction. In addition, we assembled spliceosomes on two distinct pre-messenger RNAs (pre-mRNAs; adenovirus major late, AdML, and Fushi tarazu, Ftz) so that the proteins common to both could be used as a criterion for identifying splicing-specific proteins.

Representative data for purification of AdML spliceosomes are shown in Fig. 1. AdML-M3 pre-mRNA contains three hairpins that bind to the MS2-MBP fusion protein used for affinity purification (Fig. 1a: MS2 is a bacteriophage coat protein; MBP is maltose-binding protein). AdML pre-mRNA, which lacks these hairpins, was used as a negative control. After adding the MS2-MBP fusion protein to the two pre-mRNAs, spliceosomes were assembled *in vitro*, isolated by gel filtration, affinity-selected by binding to

amylose resin, and eluted with maltose under salt conditions optimal for splicing (60 mM KCl)^{4,5}. The products of the first and second catalytic steps of splicing were detected in the gel filtration fraction for both AdML and AdML-M3 pre-mRNAs (Fig. 1b, lanes 2 and 5). In contrast, after binding to and elution from the amylose-affinity resin, only the splicing products of AdML-M3 spliceosomes were detected (lanes 3 and 6). We conclude that the spliceosomes are highly purified, as there is no detectable non-specific binding of AdML spliceosomes to the affinity resin. This conclusion is bolstered by the observation that spliceosomal small nuclear RNAs

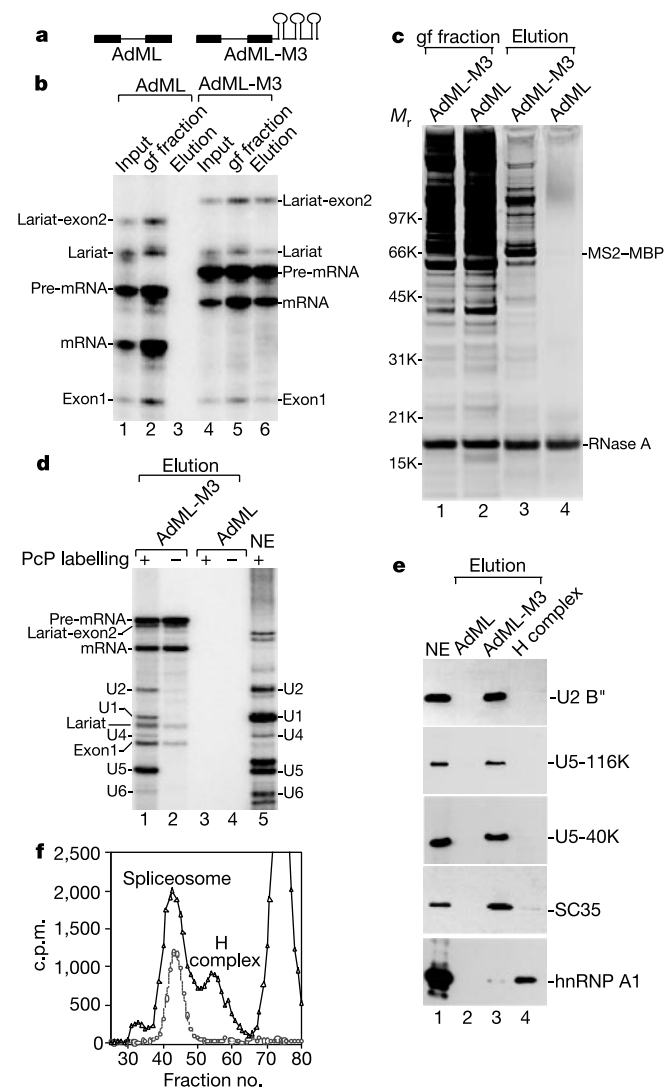


Figure 1 Isolation of spliceosomes. **a**, Schematic of pre-mRNAs. **b**, Total RNAs from the initial splicing reaction (input), gel filtration fraction (gf fraction), or final elution were separated on a 15% denaturing polyacrylamide gel. Unspliced pre-mRNA, splicing intermediates (lariat-exon2 and exon1) and products (lariat and mRNA) are indicated. **c**, Proteins from an RNase-A-treated aliquot of the gel filtration fraction or final elution were separated on 4–12% SDS-PAGE and stained with silver. The protein markers and positions of MS2-MBP and RNase A are indicated. **d**, Analysis of snRNAs. Total RNAs were extracted from the nuclear extract (NE), AdML, or AdML-M3 elution, ³²P-end-labelled (+) or not (–), and fractionated on an 8% polyacrylamide gel. Splicing intermediates, products and snRNAs are indicated. **e**, Western blot of nuclear extract (NE) and the final elutions from AdML, AdML-M3 or H complexes using the indicated antibodies. **f**, Spliceosomes were separated on a Sephacryl-S500 gel filtration column before (triangle) and after (circles) MBP-affinity purification. The positions of the spliceosome and H complex are indicated. The peak in fractions 68–80 contains degraded RNAs and free proteins.

To further examine the purity of the MBP-purified spliceosomes, we performed western blotting using antibodies that recognize representative splicing factors (Fig. 1e). As an additional control for specificity, we examined MBP-purified H complex, which contains heterogeneous nuclear ribonucleoproteins (hnRNPs) and assembles nonspecifically on RNAs in splicing extracts¹⁰. The U2 small nuclear ribonucleoprotein (snRNP) protein B'', two U5 snRNP proteins (of relative molecular mass 116,000 and 40,000 (M_r 116K and 40K, respectively))¹, and SC35, a member of the SR family of splicing factors¹¹, were detected in the AdML-M3 elution, but not in the AdML or H complex elution (Fig. 1e). In contrast, hnRNP A1 is enriched in the H complex¹⁰. These data indicate that the MBP-affinity purification is specific for AdML-M3 pre-mRNA. The data also show that the MBP-purified spliceosomes are

To determine whether the MBP-purified spliceosomes are functional, we isolated spliceosomes that had not yet undergone the first catalytic step of splicing, and used them in an *in vitro* complementation assay. The complementing extract was specifically depleted of the splicing factor SF3a⁵, which is a component of U2 snRNP and required for early spliceosome assembly^{5,8}. Thus, the depleted extracts do not support the assembly of the early spliceosome, but contain factors necessary to chase the pre-mRNA in a pre-assembled spliceosome to spliced product. To distinguish between complementation and *de novo* splicing, we used a mixture of MBP-purified spliceosomes and ‘naked’ pre-mRNA (lacking hairpins). As shown in Fig. 2a and b, splicing products were detected only from MBP-purified Ftz-M3 or AdML-M3 spliceosomes in the SF3a-depleted extract, but not from the corresponding naked

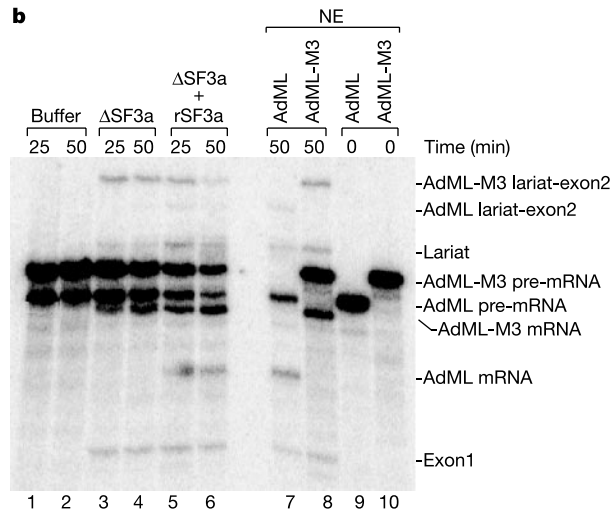


Figure 2 The MBP-purified spliceosomes are functional in an *in vitro* complementation assay. **a**, Purified early spliceosomes assembled on Ftz-M3 pre-mRNA were mixed with Ftz pre-mRNA lacking hairpins, and incubated with either buffer alone (lanes 1 and 2), with SF3a-depleted nuclear extract (lanes 3 and 4), or with SF3a-depleted extract plus recombinant SF3a protein (lanes 5 and 6) under splicing conditions for 25 min (lanes 1, 3 and 5) or 50 min (lanes 2, 4 and 6). Ftz pre-mRNA (lanes 7 and 9) or Ftz-M3 pre-mRNA

(lanes 8 and 10) were incubated under standard splicing conditions in nuclear extract for 0 or 50 min to serve as markers for the splicing products. **b**, Same as **a** except that AdML or AdML-M3 pre-mRNA was used. Total RNA was extracted and separated on an 8% (Ftz) or 16% (AdML) denaturing polyacrylamide gel. Splicing intermediates and products are indicated.

pre-mRNAs (lanes 3 and 4). We conclude that the MBP-purified spliceosomes assembled on both AdML-M3 and Ftz-M3 pre-mRNAs are functional.

The protein composition of the purified, intact and functional spliceosomes assembled on AdML-M3 and Ftz-M3 pre-mRNAs was then determined by liquid chromatography tandem mass spectrometry⁶ (LC-MS/MS). The AdML H complex was purified and sequenced for comparison. Proteins that are shared between AdML and Ftz spliceosomes are shown in Supplementary Tables I–III. (These tables contain the GenBank accession number for each protein, with live links to the protein sequence and the SWISS-PROT protein data bank. Other names for each protein are also listed.)

A total of ~145 distinct proteins were identified as shared between the two spliceosomes, including 88 known splicing factors/snRNP proteins/spliceosomal proteins (Supplementary Tables I and II). The proteins listed in Supplementary Table III were abundant in the H complex, and thus were not counted as spliceosome proteins. The proteins in Supplementary Table I are presented in functional groups, listing those that associate with the spliceosome earlier followed by those that associate later. Previous work¹ identified 43 distinct proteins in U1, U2, U4, U5 and U6 snRNPs. Of these, 41 were identified in our spliceosomes (Supplementary Table I; hLSm5 was not detected, and hLSm8 was detected only in the AdML spliceosome). The additional 47 proteins in Supplementary Table I correspond to previously known splicing factors and spliceosome components^{7–9}. Among these are all known canonical members of the SR family of splicing factors^{11–13}, all known DEXD box splicing factors⁷, all proteins found in the salt-stable spliceosome ‘core’^{2,3}, all known second catalytic step splicing factors¹⁴, and known components of the spliced messenger ribonucleoprotein (mRNP) that promotes mRNA transport to the cytoplasm¹⁵ (Supplementary Table I). A number of factors encoded by genes implicated in genetic diseases, ranging from retinitis pigmentosa (U4/U6-61K^{16,17}; U4/U6-90K¹⁸; U5-220K¹⁹) to cancer (WTAP²⁰), are also associated with the spliceosomes (Supplementary Tables I and II).

We also found 58 proteins that have not been previously identified as spliceosomal proteins in humans. These are listed in Supplementary Table II, and are organized according to the motifs in each protein. Some of these proteins (36) have previously described names and/or functions. The remainder are uncharacterized, and have been designated ‘functional spliceosome associated proteins’ (fSAPs). Among the already named proteins are 5 cyclophilin-like proteins (Supplementary Table II). Recent studies revealed a role for one cyclophilin (U4/U6-20K) in splicing²¹, and our data raise the possibility that cyclophilins (prolyl-isomerases) may have a general function in mediating conformational changes in the spliceosome. Data from several observations suggest that most, if not all, of the 58 new proteins are specifically associated with spliceosomes. First, all 58 are common to spliceosomes assembled on two different pre-mRNAs. Second, these proteins are either not detected or not enriched in the purified H complex. Third, our data (Fig. 1) show that the spliceosomes are highly purified from HeLa nuclear extract contaminants. Fourth, the new proteins are present in the spliceosome with the same relative certainty as the known splicing proteins (based on the number of unique peptides identified; compare Supplementary Tables I and II). Fifth, a large number of the proteins have domains, such as RNA recognition motifs or DEXD boxes, which are hallmarks of splicing proteins. Sixth, some of the proteins are known to co-localize with splicing factors in nuclear speckles and/or to localize in the nucleus²² (see Supplementary Table II). Last, many of the proteins (26) have known, or apparent, yeast counterparts that function in mRNA processing in *Saccharomyces cerevisiae*²³ (Supplementary Table II).

The splicing machinery is highly conserved from *S. cerevisiae* to

mammals. Among the ~145 human spliceosome-associated proteins, ~90 known or putative yeast homologues have been identified^{8,23} (Supplementary Tables I and II). The potentially greater complexity of the human spliceosome is not unexpected in light of the vastly greater complexity of splicing in metazoans compared to yeast. Indeed, most metazoan pre-mRNAs contain multiple introns, the introns are typically thousands of nucleotides, and the splicing signals are weakly conserved. Superimposed on this complexity is the high frequency of alternative splicing, which is in turn further complicated owing to regulation²⁴. Thus, many of the metazoan-specific proteins may play roles in the accurate recognition and joining of exons. The best-characterized example of such proteins is the SR family of splicing factors, which comprise at least 10 proteins in the human spliceosome, but is entirely lacking in yeast^{11–13}. SR proteins are required for both recognition of splice sites and alternative splicing, neither of which is relevant in yeast^{11–13,24}.

Splicing is thought to be coupled to several steps in gene expression, including transcription, polyadenylation and mRNA export^{25–27}. A significant fraction of the spliceosome proteome, at least 30 proteins, are either known or candidate proteins for coupling splicing to other steps in gene expression (names shown in bold font in Supplementary Tables I and II). For example, the transcription cofactor TAT-SF1 is in the spliceosome (Supplementary Table II). This protein interacts with snRNPs, and is thought to reciprocally activate transcription elongation and splicing²⁸. Additional transcription factors (for example, CA150, XAB2 and SKIP) as well as polyadenylation factors (CF I) are also in the spliceosome (Supplementary Tables I and II), and may function in coupling. We also detect Aly and UAP56 in the spliceosome, proteins that are known to couple splicing to mRNA export¹⁵ (Supplementary Table I). Aly, CA150 and SKIP were previously detected in salt-stable spliceosome cores, indicating that at least some of the coupling proteins are tightly bound to the spliceosome².

Both Aly and UAP56 are components of the conserved TREX complex, which contains proteins that function in transcription elongation in yeast²⁹. Our data reveal that every protein associated with the TREX complex is also in the functional spliceosome (Supplementary Table II), suggesting that transcription, splicing and export may all be coupled via this complex. In metazoans, most pre-mRNAs contain small exons and enormous introns. Thus, the association of the TREX complex with the spliceosome may explain how mRNA export factors are co-transcriptionally loaded onto exons, which are destined for export, and excluded from introns, which are retained in the nucleus.

Although it is possible that more spliceosomal proteins will be found, our data suggest that the purified functional spliceosomes contain virtually the complete human splicing proteome. This proteome includes all of the previously known splicing factors, snRNP proteins and spliceosomal proteins, as well as a large set of newly identified proteins. The observation that the spliceosome is associated with numerous proteins that function in coupling splicing to other steps in gene expression provides compelling evidence for the emerging concept of an extensively coupled network of gene expression machines^{25–27}. □

Methods

Plasmids encoding AdML-M3 or Ftz-M3 pre-mRNAs and the purification of spliceosomes were described previously^{4,5}. For peptide identification from database sequences using tandem mass spectrometry, the proteins in the MBP-purified complexes were fractionated from the gel origin to a distance of 3 cm into a 10% SDS–PAGE gel and stained with Coomassie blue. The gel lane was divided horizontally into ~10 slices which were individually processed to yield in-gel trypsin digestion products. The resulting peptide mixtures were separated and analysed in an automated system by nanoscale LC-MS/MS on an LCQ-DECA ion trap mass spectrometer (Thermo Finnigan) using the vented-column approach⁶. The data were searched against the non-redundant human database from NCBI with the Sequest algorithm³⁰. Peptide matches were validated by previously established criteria, including manual inspection when necessary⁶. More than 65,000 sequencing attempts (tandem mass spectra) were acquired, resulting in the identification of 8,400 unique peptides. The proteins identified are presented in

Supplementary Tables I–III. Tubulin, keratins and ribosomal proteins were found in all complexes, probably as contaminants.

Received 9 April; accepted 12 July 2002; doi:10.1038/nature01031.

- Will, C. L. & Luhrmann, R. Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell Biol.* **13**, 290–301 (2001).
- Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20**, 46–50 (1998).
- Bennett, M., Michaud, S., Kingston, J. & Reed, R. Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes Dev.* **6**, 1986–2000 (1992).
- Zhou, Z. *et al.* The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407**, 401–405 (2000).
- Das, R., Zhou, Z. & Reed, R. Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol. Cell* **5**, 779–787 (2000).
- Licklider, L. J., Thoreen, C. C., Peng, J. & Gygi, S. P. Automation of nanoscale microcapillary liquid chromatography–tandem mass spectrometry with a vented column. *Anal. Chem.* **74**, 3076–3083 (2002).
- Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).
- Burge, C. B., Tuschl, T. H. & Sharp, P. A. *The RNA World* (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 525–560 (Cold Spring Harbor Laboratory Press, New York, 1999).
- Hastings, M. L. & Krainer, A. R. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* **13**, 302–309 (2001).
- Bennett, M., Pinol-Roma, S., Staknis, D., Dreyfuss, G. & Reed, R. Differential binding of heterogeneous nuclear ribonucleoproteins to mRNA precursors prior to spliceosome assembly in vitro. *Mol. Cell Biol.* **12**, 3165–3175 (1992).
- Fu, X. D. The superfamily of arginine/serine-rich splicing factors. *RNA* **1**, 663–680 (1995).
- Tacke, R. & Manley, J. L. Determinants of SR protein specificity. *Curr. Opin. Cell Biol.* **11**, 358–362 (1999).
- Graveley, B. R. Sorting out the complexity of SR protein functions. *RNA* **6**, 1197–1211 (2000).
- Zhou, Z. & Reed, R. Human homologs of yeast prp16 and prp17 reveal conservation of the mechanism for catalytic step II of pre-mRNA splicing. *EMBO J.* **17**, 2095–2106 (1998).
- Reed, R. & Hurt, E. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* **108**, 523–531 (2002).
- Vithana, E. N. *et al.* A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). *Mol. Cell* **8**, 375–381 (2001).
- Makarova, O. V., Makarov, E. M., Liu, S., Vornlocher, H. P. & Luhrmann, R. Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6•U5 tri-snRNP formation and pre-mRNA splicing. *EMBO J.* **21**, 1148–1157 (2002).
- Chakarova, C. F. *et al.* Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa. *Hum. Mol. Genet.* **11**, 87–92 (2002).

- McKie, A. B. *et al.* Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13). *Hum. Mol. Genet.* **10**, 1555–1562 (2001).
- Little, N. A., Hastie, N. D. & Davies, R. C. Identification of WTAP, a novel Wilms' tumour 1-associating protein. *Hum. Mol. Genet.* **9**, 2231–2239 (2000).
- Horowitz, D. S., Lee, E. J., Mabon, S. A. & Misteli, T. A cyclophilin functions in pre-mRNA splicing. *EMBO J.* **21**, 470–480 (2002).
- Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
- Stevens, S. W. *et al.* Composition and functional characterization of the yeast spliceosomal pentanRNP. *Mol. Cell* **9**, 31–44 (2002).
- Cartegni, L., Chew, S. L. & Krainer, A. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285–298 (2002).
- Bentley, D. Coupling RNA polymerase II transcription with pre-mRNA processing. *Curr. Opin. Cell Biol.* **11**, 347–351 (1999).
- Hirose, Y. & Manley, J. L. RNA polymerase II and the integration of nuclear events. *Genes Dev.* **14**, 1415–1429 (2000).
- Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506 (2002).
- Fong, Y. W. & Zhou, Q. Stimulatory effect of splicing factors on transcriptional elongation. *Nature* **414**, 929–933 (2001).
- Strasser, K. *et al.* TREX is a conserved complex coupling transcription with messenger RNA export. *Nature* **417**, 304–308 (2002).
- Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements

We thank N. Dorman, E. Ibrahim and K. Magni for comments on the manuscript; C. Thoreen for database support; R. Das for SF3a-depleted nuclear extract; and R. Luhrmann and G. Dreyfuss for antibodies. HeLa cells were obtained from the National Cell Culture Center. This work was supported by the NIH (S.P.G. and R.R.).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.R. (e-mail: rreed@hms.harvard.edu).

Out of sequence

Synthesis, sequencing and other DNA-related matters

StratoSpheres Plugs

Polymer Laboratories www.polymerlabs.com

Resin to be cheerful

StratoSpheres Plugs are resin-loaded devices designed to simplify the synthesis of small molecules and peptides. The resin is entrapped so that it reacts in a normal manner, but contained in a pre-formatted, pre-packaged format eliminating the need to weigh loose resin. The composite plugs are made by co-sintering HDPE material with solid-phase synthesis resins, solid supported reagents or scavengers. StratoSpheres Plugs suit most 48- or 96-well formats and can be used in all types of agitation systems, including magnetic stirring or vortexing or with automation devices.



Cent packing: StratoSpheres plugs are small.

labelled probes. Serial amplification is used for high sensitivity, and detection reagents and protocols are optimized to maximize signal-to-noise ratios. The kits are available in two formats for detection of fluorescein labelled probes and digoxin.

3730/3730xl DNA analysers

Applied Biosystems www.applera.com

Increase productivity twofold

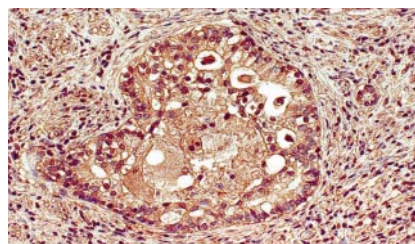
Applied Biosystems' 3730 and 3730xl DNA analysers are marketed as next generation production-scale DNA analysis systems. The 3730 analyser, a 48-capillary electrophoresis system, combines technology from the established ABI PRISM3100 genetic analyser with the latest in automation and optics, to meet the needs of high throughput. The 3730 platform can be upgraded to 3730xl specification, the 96-capillary successor to the PRISM 3700. Researchers and production facilities are expected to use the 3730xl DNA analyser for production sequencing and the 3730 for a wider range of sequencing and fragment analysis applications.

Universal ISH Detection Kits

Innogenex www.innogenex.com

Search and stain

The Universal ISH (*in situ* hybridization) Kit offers a complete non-radioactive system for localization of DNA and RNA sequences in all cells and tissues, including rodent species. The ready-to-use kit utilizes proprietary signal amplification technology for the colorimetric detection of fluorescein and digoxin



Universal ISH: seeking out RNA and DNA.

In-Fusion PCR Cloning System

BD Biosciences www.bdbiosciences.com

Rapid high-throughput PCR

The BD In-Fusion PCR Cloning System enables directional and efficient high-throughput cloning of PCR fragments. This method is compatible with any thermostable polymerase and does not require blunt end polishing or the presence of A-overhangs. In-Fusion System is claimed to generate precise, directional constructs that are ready for expression analysis with BD Biosciences' Creator Gene Cloning and Expression System.

Cloned AMV First-Strand Synthesis

Invitrogen Corporation www.invitrogen.com

Go into reverse transcription

Cloned AMV First-Strand Synthesis Kit combines the technology of Invitrogen's Cloned AMV RT with the convenience of optimized, pre-qualified reagents. Everything necessary for high-quality, routine reverse transcription is provided, including highly purified Cloned AMV RT.

All-Gone

Amresco www.amresco-inc.com

Prevent nuclease contamination

The All-Gone Glass Plate Kit offers a quick, convenient and efficient way to safeguard experiments from RNase and DNase contamination. It is a three-component kit that can eliminate nuclease contamination associated with preparing glass plates for electrophoresis experiments. NucleasEliminator Wipes are used to remove nuclease contamination from the glass plate surface. The rinse solution is for rinsing the plate off, and Acryl-Glide Wipes prevent the glass plates from sticking to acrylamide gels. The All-Gone kit contains

sufficient materials to prepare 10 slab gel sequencing plate sets or 20–40 mini gel plate sets and can be stored at room temperature.

Biochrom 30

Biochrom www.biochrom.co.uk

Manifold improvements

Hardware improvements incorporated in the Biochrom 30 amino acid analyser include new pumps with ceramic heads for dependable high-pressure operation and future expandability, and a new manifold-based valving system that improves reliability by reducing the potential for leaks. The Biochrom 30 software fully integrates instrument control and data handling functions. Coupled with a new Graphical User Interface, this is claimed to make operation intuitive and easy, allowing flexible data export and customized reporting.

Westran Clear Signal/ S PVDF

Schleicher & Schuell BioScience

www.arraying.com

Branching out

Schleicher & Schuell's PVDF membrane is optimized for proteomics, western blotting and protein sequencing applications. The previous single Westran membrane has been replaced with two new formulations, Westran Clear Signal and Westran S. With uniform pore structures, Westran Clear Signal is a 0.45 μ m pore membrane designed to produce low background and strong signal after western blotting. Westran S has increased chemical resistance, high tensile strength and a 0.2 μ m pore size, making it ideal for protein sequencing applications. Westran Clear Signal and Westran S PVDF may be ordered directly from S&S BioScience by calling 1-800-245-4024.

These notes are compiled in the Nature office from information provided by the manufacturers.

nature insight

climate & water



Cover illustration
courtesy of Taxi/Getty
Images.

The book of Genesis, one of the oldest written accounts of human thought about nature, places water second in importance only to light. In the story of the creation, day 2 already is devoted to the separation of the waters below the sky from the waters above it; day 3 begins with the separation of land from oceans.

As recognized since the writing of the Old Testament, the distribution of water within the Earth system is vitally important for humankind and life in general. Of the two most important climate parameters, temperature and precipitation, the latter simply is water and the former is dominated by water in its role of distributing heat over the planet.

A large number of feedback cycles, some of them opposite in effect to others, are involved in this distribution of the incoming energy from the Sun. Water vapour in the atmosphere is the most important greenhouse gas, but clouds and ice sheets can cool the Earth by reflecting energy back to space. Ocean currents transport heat away from the tropics and release it in high latitudes, but they also bring moisture which — under suitable conditions — can induce the growth of ice sheets. Boldly simplified, water acts as the venetian blind of our planet, as its central heating system and as the fridge, all at the same time.

How the balance between the multitude of processes will evolve in a changing climate is as important a question as it is difficult. Water in its various forms has always worked as a great amplifier of changes, such as variations in insolation or tectonic changes, that are imposed on the climate system. Similarly, water is intimately involved in the way the more recent addition of anthropogenic greenhouse gases affects climate.

The articles in this Insight explore some of these interactions between climate and the hydrologic cycle in the past, present and future.

Heike Langenberg Senior Editor

commentary:

- 188** Reducing uncertainty
about carbon dioxide
as a climate driver
L. R. Kump

review articles:

- 191** The hydrologic cycle
in deep-time
climate problems
R. T. Pierrehumbert
- 199** Links between climate
and sea levels for the
past three million years
*K. Lambeck, T. M. Esat
& E.-K. Potter*
- 207** Ocean circulation and
climate during the
past 120,000 years
S. Rahmstorf
- 215** A satellite view of aerosols
in the climate system
*Y. J. Kaufman, D. Tanré
& O. Boucher*
- 224** Constraints on future
changes in climate and
the hydrologic cycle
*M. R. Allen
& W. J. Ingram*
- 228** Towards objective
probabilistic climate
forecasting
*M. R. Allen
& D. A. Stainforth*

Editor, *Nature*: **Philip Campbell**
Insights Editor: **Karl Ziemelis**
Editorial Assistant: **Simon Gibson**
Production Editor: **Simon Gribbin**

Art Director: **Majo Xeridat**
Diagrams: **Ann Thomson**
Vicky Askew
Suzanne Coleman

Production: **Sue Gray**
Marketing: **Claire Aspinall**
Sponsorship: **Mark Greene**

Reducing uncertainty about carbon dioxide as a climate driver

Lee R. Kump

Department of Geosciences and NASA Astrobiology Institute, The Pennsylvania State University, University Park, Pennsylvania 16802, USA
(e-mail: lkump@psu.edu)

The lack of an adequate ancient analogue for future climates means that we ultimately must use and trust climate models, evaluated against modern observation and our best geologic records of warm and cold climates of the past. Armed with an elevated confidence in the models, we will then be able to make reliable predictions of the Earth's response to our risky experiment with the climate system.

The world's industrialized nations have initiated a global experiment in climate change, unwittingly at first, knowingly now, but with great uncertainty and debate about its consequences. The experiment began to have detectable consequences (in retrospect) for atmospheric composition with the clearing of Northern Hemisphere forests in the nineteenth century. The signal intensified through the twentieth century with the continued destruction of forest globally and the exponential increase in the rate of combustion of fossil fuels, the release of methane and nitrous oxide from agricultural lands and livestock, and the production and subsequent loss to the atmosphere of synthetic freons (halogenated hydrocarbons). These gases are all of consequence for climate because they are greenhouse gases: they absorb infrared energy (heat) emitted from the Earth's surface and re-emit a substantial fraction of this energy back to Earth. Their direct effects on global energy balance can be readily calculated, but the climate system they are disturbing is complex, with numerous feedback mechanisms (interdependent cause–effect relationships) that create indirect effects that are more difficult to predict.

Although it is unlikely that these feedbacks could ever reverse the climatic response from warming to cooling, the transition to new states may involve counterintuitive transients. For example, the initial warming during the last deglaciation 12,000 years ago caused a temporary return to the glacial state (the Younger Dryas; see review in this issue by Rahmstorf, pages 207–214). Policy-makers (especially in the United States) have used such uncertainties to justify delaying action to mitigate against global warming. Fortunately, several scientific approaches are being taken to reduce the uncertainties, including process studies, numerical modelling, comparison of models with the recent climate record (see the review in this issue by Allen and Ingram, pages 224–232), and investigation of ancient climates. However, recent palaeoclimate studies^{1–3} have stirred up scientific debate about the role of carbon dioxide as a climate driver and the importance of other feedback mechanisms, especially those involving water vapour, in climate change.

The role of water in the greenhouse effect

An important suite of climate processes involves atmospheric water — the amount of water vapour (itself a greenhouse gas) in the atmosphere, cloud cover and albedo

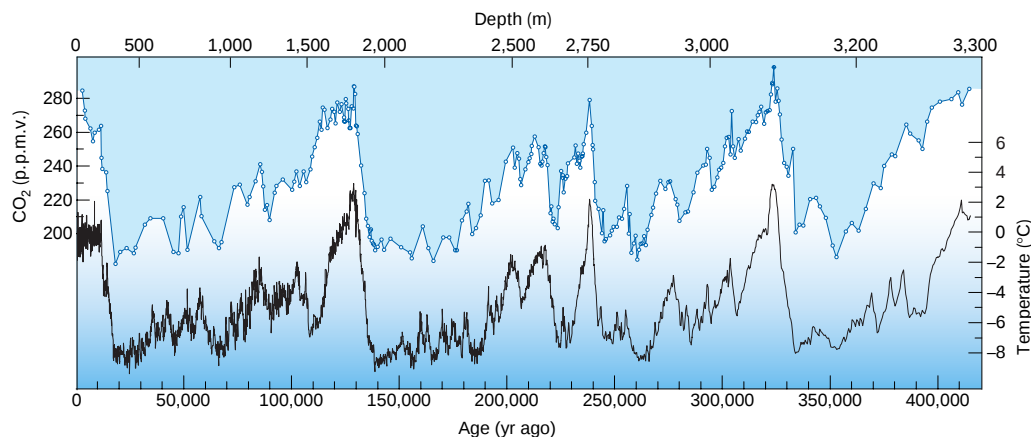
(reflectivity) of the clouds, the transport of latent heat (water vapour) through the atmosphere and the intensity and distribution of rainfall. In fact, water vapour is the most important greenhouse gas, accounting for around two-thirds of the 33°C of additional warming our planet receives because of the presence of its atmosphere. Given this, why are we concerned about the build-up of CO₂ and other greenhouse gases whose direct contributions to global warming are relatively small?

To understand the answer to this question we must distinguish 'external' versus 'internal' influences on the climate system. External influences include variations in insolation on a wide range of timescales, inputs of greenhouse gases and reflective aerosols from anthropogenic activities and natural processes such as volcanism, and, on geologic (million-year) timescales, the uplift of mountain belts and the geographical distribution of the continents. These factors are external to the climate system in the sense that although they affect climate, climate has no apparent reciprocal effect on them. (However, it has been suggested that climate may indeed influence mountain uplift and thus that tectonics is not strictly 'external' to the climate system⁴; additionally, policy changes in response to climate change may ultimately affect human activities such as fossil-fuel burning.)

External influences either perturb the climate system with short-duration disturbances (for example, injection of volcanic aerosols into the atmosphere, causing cooling; see review in this issue by Kaufman and co-workers, pages 215–223) followed by climatic response (often relaxation to the original state), or drive it with persistent forcing to new states. In this respect, not all external forcings are equal. For example, the cooling effect of sulphuric-acid aerosols produced by fossil-fuel combustion offsets some of the present warming resulting from fossil-fuel CO₂ emissions, but once the burning stops, the aerosol cooling will disappear within years to decades whereas the warming from CO₂ will last for millennia (and thus will affect other climatic factors such as the size of polar ice sheets).

In contrast, 'internal' climate influences are integral parts of the climate system, inseparably coupled through feedback mechanisms. Internal feedbacks include those that relate surface temperature to the water-vapour content and cloudiness of the atmosphere, to sea-ice coverage of the oceans, to the exchange of CO₂ between the oceans and atmosphere, and to the release of methane from wetlands. Internal feedbacks amplify or damp external forcings, but

Figure 1 The variations in atmospheric carbon dioxide concentrations and relative changes in air temperature determined from the Vostok, Antarctica ice core are tightly correlated and reveal no obvious, substantial lead-lag relationship. Figure adapted from ref. 14, with permission.



cannot in themselves drive climate change. The increase in water-vapour content of the atmosphere in response to the initial warming from elevated atmospheric CO_2 partial pressure (pCO_2) is a positive feedback that amplifies the warming. An increase in cloudiness (in particular of low clouds; see Box 1 in the review by Pierrehumbert, pages 191–198) is a negative feedback that tends to damp the warming.

It is difficult to predict the extent of amplification or damping of the external forcings by internal feedbacks. Complex general circulation models of the climate system incorporate the current understanding of the nature of these feedbacks, and predict net warming well in excess of the direct effects of CO_2 for this century (see <http://www.ipcc.ch> for a current assessment of climate-change potential by the Intergovernmental Panel on Climate Change). Much of the extra warming is predicted to come from an enhanced greenhouse effect attributable to water vapour.

Reducing uncertainties in model predictions

There is considerable uncertainty in model predictions of climate change (see the review in this issue by Allen and Ingram, pages 224–232). This results in part from the coarseness of the gridded representation of spatially continuous physical processes that numerical models must adopt, especially as it affects our ability to predict cloud cover and rainfall. The climate system has processes acting at microscopic to megascopic scales, and the fine-scale processes (for example, the formation of clouds) must be parameterized rather than treated explicitly in global models.

There is also the possibility that the models are missing key climate feedbacks, with biotic processes perhaps being neglected most in previous models. State-of-the-art climate models have added interactions with terrestrial (land-based) ecosystems. For example, current models incorporate the ability of plants to pump vast amounts of water into the atmosphere through evapotranspiration and modify the surface albedo, thus affecting local energy budgets and precipitation. Fully interactive models also allow the vegetation type to change as the climate predicted by the model changes. However, most models neglect the potentially critical role that marine algae play in the formation and reflectivity of clouds over the remote ocean⁵.

One approach to reducing model uncertainties is to evaluate how the climate system has responded to similar forcings in the geologic past. In doing so, palaeoclimatologists have discovered a number of climate paradoxes that challenge us to think in new ways about the climate system. The 400,000-year Vostok ice-core record of atmospheric pCO_2 and temperature would seem to provide the ideal demonstration that CO_2 drives climate (Fig. 1). The two records are highly positively correlated, with warm intervals occurring during

times of elevated atmospheric pCO_2 and vice versa. But cause and effect is difficult to assign. On these geologically short timescales, changes in atmospheric pCO_2 are internal to the climate system, driven by exchanges of carbon among the ocean, atmosphere and terrestrial biomass that affect and are affected by climate change. In feedback loops, the comfortable concept of cause and effect loses its meaning, and so attempts to assign them usually end in frustration. The pacemaker of Pleistocene climate change seems to be subtle changes in the Earth's orbit around the Sun (see the reviews in this issue by Rahmstorf, pages 207–214, and Lambeck *et al.*, pages 199–206), but the climatic response clearly involves CO_2 .

Climate models of the Last Glacial Maximum provide good support of the CO_2 -climate connection. When the Earth's orbital parameters are set to favour glaciation, and the maximum extent of ice sheets in the Northern Hemisphere is specified but atmospheric pCO_2 is maintained at pre-industrial levels, these models fail to achieve cooling south of the Equator. However, when reduced atmospheric pCO_2 is added to the model specifications, the tropics and Southern Hemisphere cool⁶, more in keeping with the data (in fact, such studies indicate that climate models may be underestimating the climatic sensitivity to changes in atmospheric pCO_2). Thus, despite all the uncertainties, climate models do reproduce the general features of the cooler climate of the last glacial interval, but only with reduced atmospheric pCO_2 .

We need to look beyond the glacial record, though, if we want to be able to predict the future. The range of natural variability in atmospheric pCO_2 over the past 400,000 years, from approximately 180 to 280 parts per million by volume (p.p.m.v.), does not even include the present-day human-perturbed value of 370 p.p.m.v. We thus have no analogue in the Pleistocene glacial record for climates of the future, with atmospheric pCO_2 potentially reaching 2,000 p.p.m.v. upon complete utilization of the world's coal supplies. (The known coal reserves dwarf those of oil and natural gas in terms of their ability, upon utilization, to alter the CO_2 content of the atmosphere.)

Comparison with the geologic past

Compared with much of Earth's history, we live in a time of globally cool temperatures and (probably) low atmospheric pCO_2 . A promising approach to the evaluation of climate response to greenhouse gas build-up, then, is the analysis of high-resolution palaeoclimate data sets from the more distant geologic past when warm climates prevailed. Based on these records, there are good reasons to suspect that atmospheric pCO_2 has been a primary climate driver⁷, but the evidence is not conclusive. Certain intervals of the Earth's history, such as the Middle Cretaceous (about 100 million years (Myr) ago) are characterized by fossil and geologic indicators of global warmth,

and by voluminous deposits of volcanic rocks and other indicators of abundant volcanism. Volcanoes are a chief source of CO₂ to the atmosphere, so it is reasonable to conclude that atmospheric pCO₂ was elevated during such times. However, the times of greatest volcanic activity may not correlate directly with times of greatest warmth³. Moreover, unlike the Pleistocene, there is no direct evidence for CO₂ levels in earlier times. Numerical carbon-cycle models that calculate ancient CO₂ levels, and pCO₂ proxies derived from the isotopic composition of marine organic matter or carbonate nodules in ancient soils, or from the density of stomata on fossil leaves, do generally support the relationship between climate and atmospheric pCO₂ on geologic timescales⁷.

There is also proxy evidence that greenhouse gases (CO₂ and methane) have driven more rapid climate shifts. A recent analysis of stomatal density from plant fossils deposited just after the Cretaceous/Tertiary mass extinction (65 Myr ago) suggests that atmospheric pCO₂ may have risen to 1,000s of p.p.m.v., consistent with indicators of global warmth, and implicating meteorite impact into a limestone target area⁸. Cores of marine sediments also reveal highly detailed glimpses of climate history. For example, sediments providing 1,000-year temporal resolution offer compelling evidence that an abrupt warming at the Palaeocene/Eocene boundary (55 Myr ago) was caused by a massive release of methane from sea-floor gas-hydrate deposits⁹. The timescale and magnitude of the release are similar to the projected fossil-fuel pulse, so the event could potentially be an analogue for future climate change. However, the 1,000-year resolution is limiting in this respect, and future responses are likely to differ from those of a Palaeocene lacking continental ice sheets.

Mismatches in the CO₂–climate relation

Despite these successes in linking variations in greenhouse gas concentrations to climate change in the geologic past, the oxygen isotope palaeotemperature record from 600 Myr ago to the present displays notable intervals for which inferred temperatures and pCO₂ levels are not correlated¹. One of these occurred during the early to middle Miocene (about 17 Myr ago), a time well established as a warm interval (relative to today), but with proxy evidence for low atmospheric pCO₂ (ref. 2). Moreover, whereas climate models predict tropical warming in response to elevated pCO₂, geologic data — in particularly the oxygen isotope record — indicate muted warming or even cooling at low latitudes while higher latitudes warm (the ‘cool tropics paradox’^{10–11}). A better understanding of what controls the latitudinal temperature gradient is clearly needed (see review in this issue by Pierrehumbert, pages 191–198). It is possible, however, that this paradox and the longer-term mismatch between pCO₂ and temperature proxies are artefacts, the consequence of inevitable blurring of the signal by geochemical processes that act on sediments after deposition on the sea floor^{3,12}.

Is the CO₂–climate relationship refuted by these apparent mismatches? I think not. Even if the proxies survive further scrutiny and the mismatches remain, we should not expect highly correlated records of volcanism and other external forcings, atmospheric pCO₂ and climate. In intricately coupled systems such as the climate system, with components acting on a wide range of timescales, responses of system state to stimuli can lag forcing significantly. CO₂ build-ups can lag elevated CO₂ inputs by millennia to millions of years because other processes damp the forcing; in such a case, we would expect the evidence for elevated inputs (for example, volcanic activity) to correlate instead with the rate of increase in atmospheric pCO₂. Similarly, a CO₂ increase whose radiative forcing drives a slow-changing climate component (for example, the extent of continental ice sheets during glaciation) could produce a proxy climate record (for example, ice-sheet volume as expressed in the oxygen isotopic composition of the ocean; see review in this issue by Lambeck *et al.* pages 199–206) that was poorly correlated with the CO₂ forcing.

This explanation for the out-of-phase relationship between climate and atmospheric pCO₂ may apply to the Miocene mismatch,

but is best illustrated in the ‘snowball Earth’ scenario proposed for the Late Precambrian (750–600 Myr ago; see review in this issue by Pierrehumbert). The frozen Earth accumulates CO₂ from volcanoes for millions of years until the level is sufficiently high (perhaps hundreds of times the present level) to create a greenhouse effect that can overcome the cooling effect of highly reflective continental ice sheets and global sea-ice cover. Theoretically, CO₂ is highest at the end of the snowball interval, and then drops during the warmth of the ensuing deglaciation. A similar scenario was proposed for the more modest Late Ordovician glaciation (440 Myr ago)¹³. Moreover, if thresholds must be overcome before the system can change state, then relationships become highly nonlinear and may exhibit long periods of no correlation. Switches in ocean circulation patterns and their attendant effects on climate are good examples of threshold effects (see review in this issue by Rahmstorf, pages 207–214).

Perspectives

We have made the natural world our laboratory, but the experiment is inadvertent and thus not designed to yield easily decipherable results. Consequently, we will have difficulty isolating the effects of our manipulations from natural variations in climate until the signal has risen well above the noise (and the climate perhaps has been detrimentally altered). In looking into the past history of the planet for clues to the future, we find general support for the notion that an increase in atmospheric pCO₂ will cause global warming. However, in detail the relationship is neither linear nor in phase on all timescales. Proxy indicators of global warmth do not always coincide with proxy indications of elevated pCO₂, and when they do, as in the Late Pleistocene, there is no lead–lag relationship from which one might hope to assign cause and effect. Fortunately, improved models evaluated against expanded high-fidelity palaeoclimate databases are on the horizon, and should be adequate to support policy decisions concerning the reduction of fossil-fuel CO₂ emissions. In the meantime, there are unsettling indications that these models are underestimating rather than overestimating the climatic consequences of greenhouse gas build-up. □

doi:10.1038/nature01087

1. Veizer, J., Godderis, Y. & Francois, L. M. Evidence for decoupling of atmospheric CO₂ and global climate during the Phanerozoic eon. *Nature* **408**, 698–701 (2000).
2. Pagani, M., Arthur, M. A. & Freeman, K. H. Miocene evolution of atmospheric carbon dioxide. *Paleoceanography* **14**, 273–292.
3. Wilson, P. A., Norris, R. D. & Cooper, M. J. Testing the Cretaceous greenhouse hypothesis using glassy foraminiferal calcite from the core of the Turonian tropics on Demerara Rise. *Geology* **30**, 607–610 (2002).
4. Montgomery, D. R., Balco, G. & Willett, S. D. Climate, tectonics, and the morphology of the Andes. *Geology* **29**, 579–582 (2001).
5. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).
6. Broccoli, A. J. & Manabe, S. The influence of continental ice, atmospheric CO₂, and land albedo on the climate of the last glacial maximum. *Clim. Dynam.* **1**, 87–99 (1987).
7. Crowley, T. J. & Berner, R. A. CO₂ and climate change. *Science* **292**, 870–872 (2001).
8. Beerling, D. J., Lomax, B. H., Royer, D. L., Upchurch, G. R. Jr & Kump, L. R. An atmospheric pCO₂ reconstruction across the Cretaceous–Tertiary boundary from leaf megafossils. *Proc. Natl Acad. Sci. USA* **99**, 7844–7847 (2002).
9. Dickens, G. R., O’Neil, J. R., Rea, D. K. & Owen, R. M. Dissociation of oceanic methane hydrate as a cause of the carbon–isotope excursion at the end of the Paleocene. *Paleoceanography* **10**, 965–971 (1995).
10. Barron, E. J. Eocene equator-to-pole surface ocean temperatures: a significant climate problem? *Paleoceanography* **2**, 729–739 (1987).
11. Crowley, T. J. & Zachos, J. C. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. S. & Wing, S. C.) 50–76 (Cambridge Univ. Press, Cambridge, 2000).
12. Pearson, P. N. *et al.* Warm tropical sea surface temperatures in the Late Cretaceous and Eocene epochs. *Nature* **413**, 481–487 (2001).
13. Kump, L. R. *et al.* A weathering hypothesis for glaciation at high atmospheric pCO₂ in the Late Ordovician. *Paleoclimatol. Palaeogeogr.* **152**, 173–187 (1999).
14. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).

Acknowledgements

I thank R. Alley, M. Arthur and R. Pierrehumbert for their constructive criticism of drafts of this commentary. Research on ancient environments by L.R.K. at Penn State is supported by grants from NSF Biocomplexity, Geology and Paleontology, and NASA Astrobiology programmes.

The hydrologic cycle in deep-time climate problems

Raymond T. Pierrehumbert

Department of the Geophysical Sciences, The University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA
(e-mail: rtp1@geosci.uchicago.edu)

Hydrology refers to the whole panoply of effects the water molecule has on climate and on the land surface during its journey there and back again between ocean and atmosphere. On its way, it is cycled through vapour, cloud water, snow, sea ice and glacier ice, as well as acting as a catalyst for silicate-carbonate weathering reactions governing atmospheric carbon dioxide. Because carbon dioxide affects the hydrologic cycle through temperature, climate is a *pas des deux* between carbon dioxide and water, with important guest appearances by surface ice cover.

When sunlight evaporates water from the ocean surface, the energy is stored in the form of latent heat in atmospheric water vapour. The flow of this fluid sunshine to distant places, where it releases its energy by condensation, forms one of the main modes of energy transport in the atmosphere. Water also affects the radiation balance of the planet through the water vapour greenhouse effect and cloud greenhouse effect, and through reflection of sunlight by clouds and ice. The rate of precipitation of water from the atmosphere determines the rate of chemical weathering of silicate rocks, and hence ultimately the carbon dioxide content of the atmosphere. In the following review, I will discuss in turn the way these three aspects of the hydrologic cycle helped to make past climates of the Earth so different from the present one.

The 'snowball Earth' concept provides stimulus for thinking about the hydrologic cycle in climates radically different from the present. Snowball Earth refers to a situation in which the ocean becomes ice-covered all the way from the poles to the tropics, possibly with the accompaniment of near-total land coverage by glaciers. There is evidence that this happened at least twice during the Neoproterozoic, about 600–700 million years ago^{1,2}.

In discussing water and energy transport, I focus principally on differences in pole-to-equator temperature gradients in warm and cold climates. An interplay of latent heat scaling, temperature gradient and eddy activity conspires to keep net north-south atmospheric energy transport relatively unchanged between Last Glacial Maximum (LGM) conditions and Cretaceous-style hothouse conditions, although the constancy is expected to break down both in ultra-cold snowball conditions and in ultra-warm conditions prevailing in the aftermath of deglaciating a snowball.

My discussion of the radiative effects of water vapour focuses on cloud, water vapour and ice-albedo feedbacks as they affect the conditions for initiation and termination of the snowball state (for a broader review of radiative water vapour feedbacks, see refs 3, 4). The principal conclusion here is that most questions one might wish to address regarding snowball Earth are exquisitely and distressingly sensitive to the details of cloud behaviour.

Finally, in discussing the weathering effects of water, I emphasize the way energetic considerations constrain the precipitation-temperature relation. Low precipitation (and

hence slow weathering) is unambiguously and unsurprisingly expected in cold conditions, but in sufficiently warm conditions energetic limits come into play that significantly affect the operation of the silicate weathering thermostat.

Water vapour and energy transport

The answer to virtually every grand-challenge climate problem hinges on the behaviour of the pole-to-equator temperature gradient ΔT , or equivalently, the meridional heat transport carried by the atmosphere and ocean. The latent heat content of water vapour can have a profound effect on the efficiency with which heat is transported.

ΔT affects the initiation of ice ages, with high ΔT making it easier to grow glaciers and sea ice at high latitudes. But in very cold, low- CO_2 climates, a low ΔT makes it easier for ice to penetrate into the tropics and trigger a global glaciation. The threshold CO_2 needed to deglaciade a snowball state depends critically on ΔT , as a large value makes it easier to melt tropical ice, whereafter ice-albedo feedback takes care of the rest of the planet. A small value of ΔT helps to explain the Cretaceous and similar climates, as it is then possible to keep the poles ice-free with moderate elevations of greenhouse gas concentrations, without violating observational constraints on tropical temperatures.

It has long been noted that warm climates like that of the Cretaceous tend to have smaller ΔT than the present, whereas cold climates like that of the LGM have much larger ΔT , a phenomenon known as 'polar amplification'. Polar amplification is also apparent in projections of global warming, in that doubled- CO_2 experiments show greater warming at the poles than in the tropics. This overall picture of climate change has been damaged by recent revisions to estimates of LGM and Cretaceous tropical temperature, but not destroyed. On the cold side, LGM Greenland temperatures⁵, together with recent estimates of tropical cooling⁶, suggest that the surface temperature difference between 75 °N and the tropics increased to 50 K from its modern annual-mean value of about 33 K. On the warm side, deep-sea Eocene and Early Palaeocene temperatures⁷ suggest polar temperatures of 283–287 K, with similar (or even warmer) values prevailing in the Cretaceous⁸. Recently, it has been argued that the tropics warmed much more during the Cretaceous than previously believed, potentially up to 4 K warmer than the present tropics⁹. Even the hot-tropics scenario yields a ΔT of 19–23 K, which is considerably lower

than the present value, although not so much so as the more conventional cool-tropics scenario. Can changes in latent heat flux account for polar amplification?

The meridional fluxes of dry and moist energy are:

$$F_d(\phi) = \langle \rho c_p v T + \rho g v Z \rangle, \quad F_w(\phi) = \langle \rho v L q \rangle \quad (1)$$

where F_d and F_w are the fluxes of dry static energy and latent heat, ϕ is latitude, ρ is air density (including water vapour), c_p is the mean specific heat of air (again including water vapour), v is meridional velocity, Z is geopotential height, L is latent heat of condensation and q is the specific humidity (that is, the proportion of water vapour to the total mass of the air). The dry static energy is dominated by vT in the mid-latitudes, but the geopotential term becomes important in the tropics. Angle brackets denote an integral over altitude and an average along the latitude circle. Figure 1 shows an estimate of the annual-mean dry and moist fluxes for the years 1979 to 1995, recomputed from the data described in ref. 10. In the tropics, the latent heat flux is actually towards the Equator, owing to the transport of moist air in the low-level branch of the Hadley circulation. The equatorward latent heat transport is more than cancelled by the poleward flux of dry static energy. Net tropical flux is made poleward by the vZ term, characterizing the warming of air due to compression in the sinking branch of the circulation. Moisture actually makes it harder for the Hadley cell to transport heat out of the deep tropics, because the same circulation that warms the subtropics by compressional heating and cools the deep tropics by adiabatic expansion of rising air, of necessity also carries latent heat from the subtropics into the deep tropics. In the mid-latitudes, the latent heat flux is poleward, attaining a maximum value of about 2 petawatts (PW, or 10^{15} W), which is approximately half of the total flux.

How might this situation change in a much warmer or much colder climate? In the mid-latitudes, the net flux can be estimated by

$$F_d + F_w \sim \langle \rho \rangle V c_p \left(\delta T + \frac{L \delta q}{c_p} \right) \quad (2)$$

where V is the typical meridional velocity scale, δT is the typical temperature fluctuation (assumed to be proportional to ΔT) and δq is the typical specific humidity fluctuation. Moisture is picked up at the subtropical ocean surface, which is nearly saturated. Because the heat-transporting atmospheric trajectories go poleward and upward, the relevant moisture contrast that the mixing acts upon is not between the high- and low-latitude values of boundary-layer specific humidity, but rather between the low-latitude values and the values aloft, which are essentially zero by virtue of the low temperatures there. Thus, I estimate δq as $q_s(T)$, where T is the subtropical temperature and q_s is the saturation specific humidity. Now, one can define the equivalent temperature $T^* = L \delta q_s / c_p$, which is the size of temperature fluctuation that would yield a dry heat transport equivalent to that due to moisture.

A graph of T^* is given in Fig. 2. At high temperature, the atmosphere is all steam, $q_s \rightarrow 1$ and $T^* \rightarrow 1,250$ K. In this limit, transport is wholly dominated by moisture. Moreover, a very weak wind suffices to produce an enormous heat flux. At temperatures of around 290 K, the scaling correctly predicts dry and moist heat fluxes to be equally important. As the slope is large at these temperatures, the importance of moisture changes greatly even for modest changes in T , on the order of 4 K. Latent heat transport is still an important factor for temperatures near freezing, but T^* falls to 1 K at 250 K, at which point latent heat transports are essentially negligible. Completing the estimate of heat transport requires a theory of how V changes as ΔT and moisture change. This unfortunately is beyond the state of the art, and so I confine further discussion to an examination of latent heat flux changes in a range of climate simulations.

The LGM provides an excellent laboratory for studying the behaviour of heat fluxes in a colder climate. In view of the considerable increase in ΔT at the LGM, one might expect greatly elevated heat

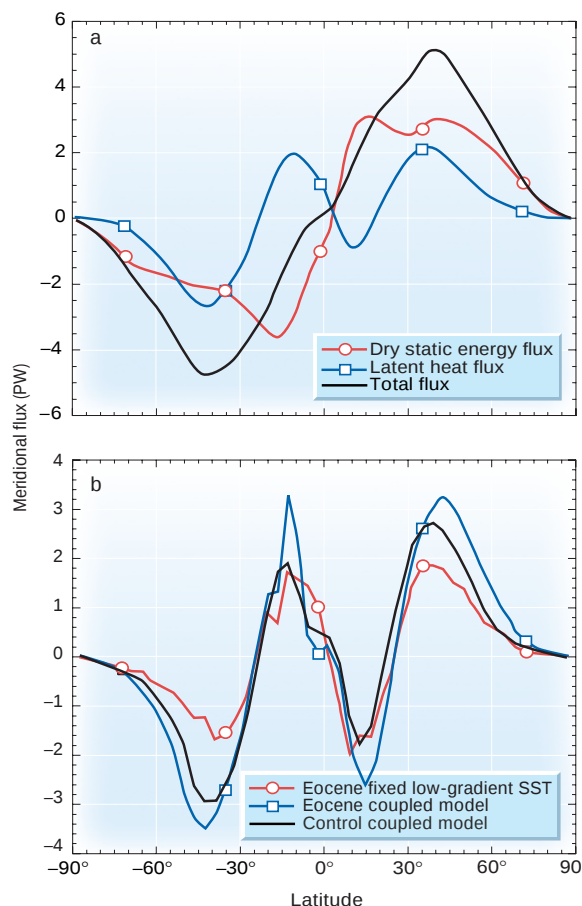


Figure 1 Northward heat flux in the atmosphere, broken down into moist and dry contributions. **a**, Observed annual-mean atmospheric energy transports computed from analyses made by the National Centers for Environmental Prediction. **b**, Latent heat flux from simulations of pre-industrial modern climate (control coupled model) and two hypothetical Eocene climates (simulations carried out with fixed low-gradient SST or with a coupled atmosphere–ocean model). Heat flux expressed in petawatts (1 PW = 10^{15} W).

fluxes. Increasing ΔT makes δT bigger in equation (2) and also makes V bigger, as an increased gradient provides a greater energy supply for synoptic eddies. In fact, simulations show little increase in the heat exported from the tropics by atmospheric fluxes^{11,12}. This is largely because a drop in latent heat transport associated with cooler subtropics compensates the increase in dry static energy transport, but there are also subtle dynamical issues involved¹³. An important consequence of the relative constancy of the heat export is that introduction of Northern Hemisphere ice sheets alone does little to cool the tropics; one must instead invoke CO_2 reductions and oceanic effects — perhaps amplified by cloud and water vapour feedbacks — to get the deep tropics to cool much. Energy-balance models that simply diffuse temperature will seriously misrepresent fluxes in such climates. Incorporating a hydrologic cycle by diffusing moisture horizontally fails to take into account the importance of lifting and cooling of air parcels. However, more sophisticated idealized models incorporating more dynamical processes¹⁴ show considerable promise.

The hard-snowball state provides an example of a yet colder climate. For a frozen Earth, there is little thermal inertia so the temperature pattern responds instantaneously to the seasonal cycle of insolation. In winter, polar temperatures fall to 190 K and the tropics are a chilly 240 K (see ref. 15; the mixed-layer simulation discussed in

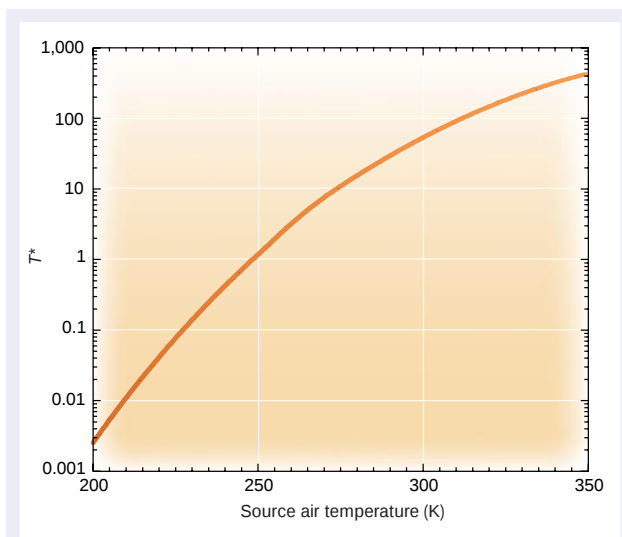


Figure 2 Characteristic temperature determining the relative importance of latent heat flux as a function of surface temperature. $T^*(T)$ is the magnitude of temperature fluctuation needed for dry heat transport to equal latent heat transport by water vapour. The graph shows the increasing dominance of latent heat flux as temperature increases.

ref. 16 yields similar values). In such a cold climate, there would be little or no latent heat flux, which is probably why such a large gradient can be maintained despite the fact that the weak solar absorption by ice means that little heat needs to be transported to even out the gradient. As CO_2 increases and the tropics warm towards freezing point, latent heat should become significant. It is an open question, though, whether the temperature gradient decreases as deglaciation is approached.

Atmospheric heat fluxes cannot maintain the low ΔT believed to have prevailed in the Cretaceous, and it has been conjectured that ocean heat transports could supply the missing flux¹⁷. But recent simulations from coupled atmosphere–ocean models do not support this supposition. In a model generated by the Geophysical Fluid Dynamics Laboratory in Princeton, permanent polar ice can be eliminated with a CO_2 level of 1,200 p.p.m., but under these conditions the tropics also warm considerably, reaching zonal mean temperatures of 302 K, or about 5 K warmer than the present tropics¹⁸.

Transport statistics are not available for this model run, so instead I present computations based on a series of Eocene simulations carried out with a climate system model developed by the National Center for Atmospheric Research (see refs 19, 20 for a description of the simulations). Latent heat fluxes for the Eocene are shown in Fig. 1b. The simulation labelled ‘Eocene fixed’ was carried out with fixed low-gradient sea surface temperature (SST) with $\Delta T = 17$ K and tropical temperatures of 300 K, conditions that suppress polar ice. Case ‘Eocene coupled’ was carried out with a coupled atmosphere–ocean model with Eocene geography and an elevated CO_2 level of 560 p.p.m. For comparison, a modern pre-industrial simulation with CO_2 at 280 p.p.m. is included (case ‘Control coupled’); the flux for this case is similar to that yielded in higher-resolution simulations driven by observed modern SST (not shown). The Eocene coupled simulation leads to warming in both the tropics and the polar regions as compared to the control, but not enough to eliminate polar ice completely. The warming is 2 K in the tropics, but 7 K at latitudes 50°N and 50°S where ice has been melted. There is thus a modest reduction in temperature gradient out to these latitudes.

In the Eocene-fixed simulation, the lower gradient reduces V and the greater moisture content of subtropical and mid-latitude air is not able to compensate; as a consequence, the flux drops considerably in comparison to present-day fluxes. In the Eocene-coupled case, the

ocean fluxes are not sufficient to maintain a low gradient, so the tropical temperature warms considerably. Although this increase does lead to a modest increase of 0.5 PW over the control in peak mid-latitude latent heat flux, it is insufficient to substantially affect the gradient.

Going into cold climates, the attenuation of the hydrologic cycle has the expected effect of reducing latent heat fluxes and allowing a larger pole-to-equator temperature gradient than would otherwise be possible. Going into warm climates, the increase in moisture content is offset by a reduction in eddy activity, and thus the heat transport changes little. Enhanced polar warming occurs in those regions where ice has disappeared, but the temperature gradient is not further moderated by latent heat feedbacks. The behaviour of latent heat flux makes it harder rather than easier to maintain a low-gradient Cretaceous or Eocene world. Current observational developments⁹ are contributing towards a better understanding of how low a gradient really needs to be explained, although this intriguing development must be regarded as tentative until it is confirmed by more (and preferably farther offshore) data. Future refinements of ocean dynamics or of radiative mechanisms²¹ offer some hope for reduction of the gradients in coupled simulations. The chasm has not yet been bridged, but it no longer seems quite so wide as it once did.

Radiative effects of water

The chief questions regarding snowball Earth are: How could Earth enter the snowball state and, once there, how could it get out? The latter is particularly important to the viability of the snowball hypothesis, as it is indisputable that if the Earth indeed were ever in a snowball state, it eventually deglaciated. Scenarios for entering the snowball state invariably posit attainment of anomalously low CO_2 concentrations, helped along by the 6% fainter Sun prevailing in the Neoproterozoic. Exit strategies conversely call on greatly elevated CO_2 concentrations; estimates based on thickness of cap carbonates or the likely amount of CO_2 degassed during the glaciated period would be consistent with deglaciation at CO_2 partial pressures in the vicinity of 0.2 bar. At present, it is impossible to rule out the possibility of much higher CO_2 concentrations at deglaciation, but the weak logarithmic dependence of the outgoing long-wave radiation (OLR) on CO_2 means that doubling or tripling the estimate of maximum allowable CO_2 at deglaciation does not greatly alter the picture presented here. The radiative properties of water vapour and clouds enter crucially in both the glaciation and deglaciation problems.

There have been several studies of initiation of the snowball state using general circulation models (GCMs) coupled to a mixed-layer ocean without ocean dynamics^{15,22,23}. Some succeed in triggering a global glaciation when CO_2 is reduced to 100 p.p.m., and others do not, with no clear picture emerging on the controlling factors. The use of specified ocean heat transports estimated from the present ocean in such models is problematic. Certainly, if ocean heat transports are eliminated altogether, the Earth falls easily into a snowball state at low CO_2 , although dynamic ocean heat transport has been found to be a potent inhibitor of glaciation¹⁶. The coupled-model results in ref. 16 should not, however, be construed as ruling out the possibility of global glaciation, in view of the relatively short simulations used, the highly idealized sea-ice model, and myriad other uncertainties plaguing ocean simulation.

There has been less study of the deglaciation problem. Deglaciation studies have been limited to idealized energy balance models²⁴, as current GCMs do not generally take into account all absorption bands that become important at high CO_2 , and often assume that CO_2 is a trace gas for the purpose of convective and thermodynamic computation. But the sensitivity of global glaciation and deglaciation to water vapour and clouds may present a severe impediment to obtaining conclusive results from either theoretical or modelling studies.

I will use an idealized model to quantify the importance of cloud and water vapour effects. The temperature of a solar-heated planet is determined by the amount of energy received from the Sun and the

temperature required to allow the planet to lose the required amount of energy to space. In the simplest form of climate model, I characterize the entire planet by a single surface temperature T , which one may think of as the global-mean temperature. The temperature is then determined by solving

$$(1 - \alpha(T))L/4 = I_+(T, p\text{CO}_2, \text{RH}, \text{clouds}) \quad (3)$$

where α is the planetary albedo (dependent on T via temperature dependence of cloud and ice cover), L is the solar constant at the planet's orbit, and I_+ is the outgoing infrared radiation flux exiting

the top of the atmosphere. For a given temperature, I_+ depends on the concentration $p\text{CO}_2$, the atmospheric relative humidity RH, and an array of cloud parameters. For clear skies, I computed $I_+(T, p\text{CO}_2, \text{RH})$ using a radiation model²⁵, which can accurately handle arbitrarily high CO_2 . The vertical profile of temperature is assumed to be the moist adiabat proceeding from the specified surface temperature T , patched to an isothermal stratosphere. Relative humidity is assumed constant in height, at a stipulated value. The assumption of fixed relative humidity implies that the water vapour content of each layer of the atmosphere increases with temperature in proportion to the local water-holding capacity, expressed by the Clausius–Clapeyron

Box 1

Climate sensitivity and the dependence of outgoing long-wave radiation on temperature

A planet can lose energy to outer space only through emission of electromagnetic radiation. For the range of planetary temperatures typically considered, this radiation is in the infrared range, and the flux of the outgoing radiation is known as outgoing long-wave radiation (OLR). The temperature of a solar-heated planet settles at the value that allows the OLR to balance the absorbed solar radiation; hence the shape of the dependence of OLR on temperature holds the key to the most fundamental climate phenomena. For an atmosphere-free planet having a uniform surface temperature T , the OLR would satisfy the Stefan–Boltzmann law, σT^4 . Atmospheres modify the OLR by allowing the planet to radiate at temperatures colder than the planetary surface. The consequent decrease in OLR is a direct consequence of the fact that temperature decreases with height, and thus the strength of the greenhouse effect is sensitive to the lapse rate.

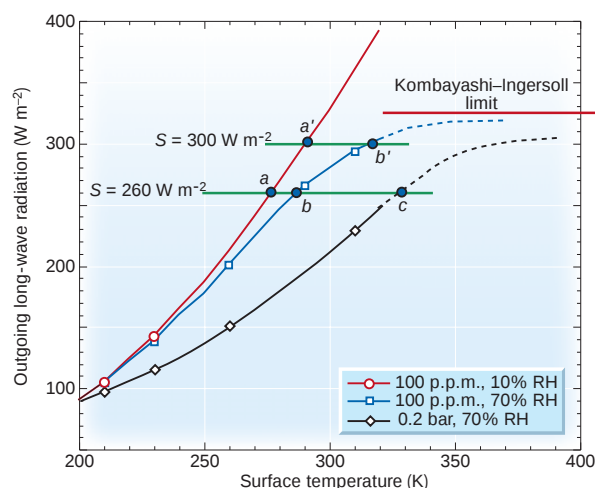
The figure opposite shows OLR as a function of surface temperature, computed under the following assumptions: (1) the vertical temperature profile is given by the moist adiabat, patched to a 200-K isothermal stratosphere, and (2) the specific humidity at each height is a fixed fraction of the saturation value given by the temperature at that height. The latter assumption is known as the ‘fixed relative-humidity assumption’. Once these assumptions are made, the OLR can be computed as a function of the surface temperature using any standard radiative-transfer model (the curves here were computed using the model of ref. 25).

Addition of a greenhouse gas to an atmosphere reduces the OLR for fixed T , with the result that temperature must increase in order to bring the radiation budget back into balance. For absorbed solar radiation $S = 260 \text{ W m}^{-2}$, 100 p.p.m. CO_2 and 10% relative humidity, the temperature is 276 K (point a in the figure). If the relative humidity is increased to 70%, the temperature increases to 288 K (point b). If the greenhouse forcing is further augmented by massively increasing the amount of CO_2 to 0.2 bar, the temperature increases to ~330 K (point c).

Because the concentration of water vapour (a greenhouse gas) increases with temperature, it acts to reduce the slope of the OLR curve. This increases the sensitivity of climate to changes in radiative forcing, as is illustrated in the figure by increasing solar absorption to 300 W m^{-2} . With 10% relative humidity, the warming from point a to a' is 14 K, but with 70% relative humidity, the warming from point b to b' is 30 K. This is known as the water vapour feedback. The ultimate water vapour feedback is the runaway greenhouse, manifest in the fact that the flattening of the OLR curve at high temperatures leads it to asymptote to a value known as the Komabayashi–Ingersoll limit. This is the limit to how fast a planet with a moist atmosphere can lose energy by infrared radiation. For a saturated atmosphere, the limit depends primarily on the gravity of the planet, which determines the mass of atmosphere contained above a given pressure level. The precise value one obtains for the limit depends on the treatment of infrared absorption properties of water vapour at high temperature and pressure. In the figure above, we have sketched in the value from ref. 42 (about 320 W m^{-2}), and the reader is referred there for further details.

When the solar forcing exceeds the Komabayashi–Ingersoll limit, the temperature continues to increase until the surface temperature is several thousand degrees, at which point all the ocean has been evaporated into the atmosphere. This is followed by disassociation of water into hydrogen and oxygen, after which the light hydrogen escapes to space making the water loss irreversible. Once there is no precipitating water reaching the ground, it becomes impossible to bind up atmospheric carbon dioxide in carbonate rocks. This scenario, or some version of it, is the prevailing view of how Venus was consigned to a climate so much less habitable than Earth's. The Komabayashi–Ingersoll limit is also relevant to impact heating of planets during their early stages, as it determines how quickly a planet can condense out the transient post-impact steam atmosphere.

The assumption of fixed relative humidity is a convenient starting point for simple models, but as the atmosphere aloft is highly undersaturated, it is dubious to consider humidity to be strictly a function of local temperature. Atmospheric humidity is in reality governed by a bewildering array of dynamical and microphysical processes for which useful idealized models are not yet available. Clouds pose an even greater challenge. The picture of the runaway greenhouse given above was based on clear-sky reasoning, but a saturated atmosphere would be thick with clouds. These clouds would block nearly all solar radiation, but would also be very effective at limiting OLR. The resulting climate would be determined by the delicate balance between a trickle of incoming solar radiation and a trickle of OLR, much as has been proposed to result from CO_2 clouds on early Mars. There have been a few tentative steps towards including clouds in the runaway greenhouse calculation⁴², but the cloud physics regime in which the cloud-forming substance is the primary component of the atmosphere is a very unfamiliar one, and offers many exciting prospects for future work.



relation. This is far from inevitable, as air aloft is highly undersaturated. However, the fixed relative humidity *ansatz* does seem to capture the radiative behaviour of water vapour in full GCMs³. Most climate phenomena, including water vapour feedback, can be understood in terms of the shape of $I_{\lambda}(T)$, also known as OLR (see Box 1).

Following common practice, the surface albedo $\alpha_s(T)$ is fixed at the value α_i for ice for temperatures $T < T_i$ where the planet is completely frozen, and at the smaller ice-free value α_o for $T > T_o$ when the planet is assumed ice-free. For $T_i < T < T_o$ one linearly interpolates between the limiting values. The presumed pole-to-equator temperature difference enters into the choice of T_i and T_o , as both temperatures would be equal to the freezing point of water for an isothermal planet, but the two values separate as ΔT increases. The cloudy-sky planetary albedo is then determined as a function of α_s and the cloud albedo α_c , using equation (1) in ref. 26. The albedos do not simply add, because sunlight reflected by the clouds does not have the opportunity to be further reflected by the surface. Thus, clouds over ice have a weaker effect on absorbed solar radiation than do clouds over ocean. In contrast, clouds have nearly the same effect on reducing OLR regardless of the underlying surface. As a result, the net cloud effect is a weak cooling over low-albedo surfaces, but adds up to a significant warming over ice. Because of this asymmetry, clouds profoundly affect initiation and termination thresholds.

Cloud albedo was parameterized as a function of T by assuming cloud water path to be a fixed fraction (10%) of saturation humidity at each height, and then calculating the top-of-atmosphere albedo using a radiative-transfer model²⁷. In this parameterization, the cloud albedo vanishes at very low temperatures where there is little water in the air; however, the albedo remains significant even near the freezing point, as it takes rather little water to make a radiatively significant cloud. To determine cloudy-sky OLR, I assumed a fixed mean cloud-top temperature, and assumed that the cloud tops are high enough to control the OLR without reference to the greenhouse gases in the overlying atmosphere. This is equivalent to fixing the cloudy-sky OLR at a stipulated value OLR_{cloud} . I have tested this assumption against calculations with a full radiation model in which cloud water content is assumed proportional to saturation humidity, and found it to be accurate. From the standpoint of OLR, clouds act much like water vapour except that, kilogram for kilogram, their infrared activity is much more potent. Increasing the atmospheric temperature increases the mass of cloud water, causing the atmosphere to radiate from higher altitudes, with the result that the effective radiating temperature changes little with surface temperature — much as in the runaway greenhouse (Box 1). In this cloud parameterization, water vapour itself has no effect on OLR.

For both OLR and albedo, the radiation balance is completed by blending the clear-sky and cloudy values according to a specified cloud fraction. Additionally, the albedo is adjusted so as to allow for solar absorption by the atmosphere. Of course, clouds are not just a function of temperature. They are strongly affected by microphysics and dynamics, among other things. For example, there are few high

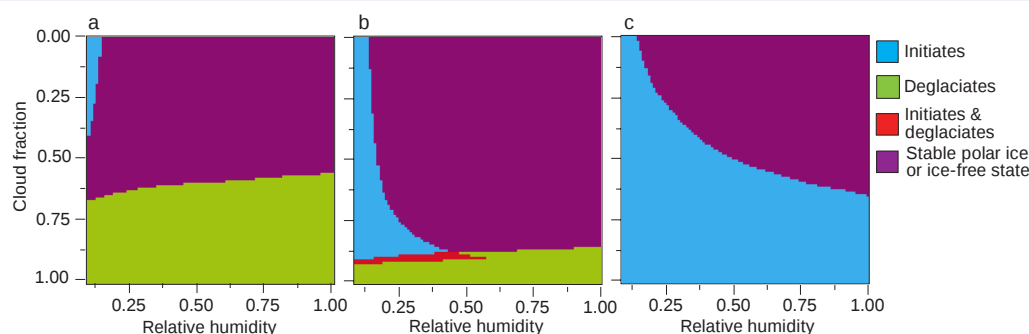
clouds in the subsiding subtropical regions, despite the warm temperatures there. In the idealized model, all such effects are wrapped into the effective cloud fraction, which will be varied as a free parameter. Feedbacks arising from systematic changes in cloud fraction as the climate state changes cannot be treated in this formulation, but one can still gain an appreciation of the kind of climate that accompanies a given amount of cloudiness.

Unless CO_2 concentration is extremely high, a fully glaciated Earth reflects enough sunlight to remain frozen until CO_2 increases enough to trigger deglaciation. This is true even for present conditions. The problem is not maintaining the snowball state, but rather accounting for how the system can be forced into that state from a state with ice cover (if any) limited to high latitudes. Thus I consider the system to initiate a snowball only when the stable ice-free or partially ice-covered state ceases to exist, in which case the planet is forced to enter a snowball state regardless of the initial conditions.

Figure 3 shows regime diagrams indicating where, as a function of cloud fraction and clear-sky relative humidity, the snowball state initiates at 100 p.p.m. CO_2 , and where the state deglaciates when CO_2 is raised to 0.2 bar. These choices are somewhat arbitrary, but serve as a reasonable basis for probing the sensitivity of the system. When $OLR_{cloud} = 120 \text{ W m}^{-2}$, the snowball state is initiated when there is sufficient cloud cover, with less cloud cover needed at low relative humidity (note that, in these calculations, relative humidity refers to the humidity of the clear-sky areas between clouds). However, there is no combination of cloud cover and humidity that allows the planet to deglaciate. When $OLR_{cloud} = 90 \text{ W m}^{-2}$, the clouds have greater warming effect and the planet can deglaciate. Under these conditions, however, the planet does not enter the snowball state unless humidity and cloud cover are severely reduced. When $OLR_{cloud} = 100 \text{ W m}^{-2}$, there is a tiny sliver of parameter space in which the planet can both initiate and deglaciate a snowball, but initiation requires dry, clear-sky conditions. Note that there is considerable sensitivity to clear-sky humidity even (in fact, especially) under dry conditions. Small amounts of water vapour can make a big difference to climate, because the radiative effect of water vapour is logarithmic in the water vapour concentration^{3,4}.

The results are consistent with GCM studies which show that (at least in the absence of dynamic ocean heat fluxes) it is possible to initiate a snowball state even in the face of cloud feedbacks, and that cloud feedback can actually help the initiation. But it is particularly unsettling that changes in the cloud radiative properties by less than 30 W m^{-2} can completely change the regime outcome; this number is well within the uncertainty of cloud parameterizations even for the well-observed modern climate. It comes as no surprise, then, that GCMs should differ so much in their initiation behaviour. To be fair, it must be said that clouds are not the only, or even chief, source of uncertainty. Small changes in the ice albedo, for example, can also have significant consequences; indeed, the difficulty in achieving globally glaciated states in the simulations of Chandler and Sohl²³ may have been due at least in part to the use of an unrealistically low

Figure 3 Regime diagrams showing effects of cloud cover and clear-sky relative humidity on initiation and termination of the snowball Earth state. All calculations were carried out with Neoproterozoic insolation, and an ice albedo of 0.7. **a**, $OLR_{cloud} = 90 \text{ W m}^{-2}$; **b**, $OLR_{cloud} = 100 \text{ W m}^{-2}$; **c**, $OLR_{cloud} = 120 \text{ W m}^{-2}$.



ice albedo. There is no one single, true value for ice albedo, and the range of surface types that need to be considered offers interesting and novel possibilities for additional radiative feedbacks²⁸.

Another implication of these results is that termination of the snowball state cannot occur at plausible CO₂ levels without considerable help from cloud effects. It is encouraging to see that there are at least plausible settings of fundamental optical cloud properties that allow the snowball state to terminate, as previous estimates of the termination threshold fixed the net cloud radiative forcing at its present value²⁴ — clearly a perilous assumption in view of the disparate influence of clouds over ice compared with clouds over water.

The snowball Earth problem has the paradoxical implication that it is now necessary to understand very warm climates — much warmer than the Cretaceous — which are nonetheless short of the conditions for a runaway greenhouse. Such climates occur in the aftermath of the snowball state, when the planet has deglaciated but CO₂ levels are still at the high levels attained during the frozen state. Based on the idealized model with 50% cloud cover and 50% relative humidity, and OLR_{cloud} = 100 W m⁻², the mean temperature rises to 332 K in the aftermath; at 100% relative humidity it would be 345 K. Happily the deglaciation does not trigger a runaway greenhouse, which would be catastrophic for both life on Earth and the viability of the snowball hypothesis. The tropical temperatures that accompany these mean temperatures depend on the efficiency of horizontal heat transport. From Fig. 2, it is clear that transports would be dominated by latent heat, but the temperature gradient nonetheless remains an open question, because it is unknown how the intensity of heat-transporting eddies would scale in such a warm climate.

A further challenge to modelling the post-snowball climate is that water vapour makes up over 10% of the total air pressure, so that precipitation has a marked effect on surface pressure, and hence dynamics. The consequences of this feedback are not invariably incorporated in current circulation models, and represent very unfamiliar territory, but some pioneering work in this direction can be found in ref. 29.

Precipitation–temperature relation and weathering

Four billion years ago, the Sun was about 30% fainter than it is at present, and yet the Earth did not spend its first three billion years of life in an unrelenting snowball state. This is known as the ‘faint young Sun’ problem³⁰. The currently favoured resolution is that CO₂ concentration was much higher earlier in the Earth’s history, and that feedback between temperature and weathering of silicate rocks into carbonate controlled CO₂ levels so as to maintain an equable climate (ref. 31; and see refs 32, 33 for a discussion of the earlier literature as well as current thinking on the subject).

In the prevailing view, temperature affects chemical weathering primarily through its effect on precipitation. It is thus the presumed increase of precipitation with temperature that leads to the stabilizing CO₂ thermostat. Although GCMs do yield increased precipitation in response to modest increases in temperature, it is not generally recognized that there are energetic constraints limiting the slope of the precipitation–temperature relation. If the temperature difference δT_s between the boundary-layer air and the surface is not too large, the surface energy budget can be linearized to yield

$$S_{\text{surf}} = E + F = (E_o(T) + a(T)\delta T_s) + (F_o + b(T)\delta T_s) \quad (4)$$

where S_{surf} is the surface-absorbed solar radiation, E is the evaporative heat flux, F is the sum of radiative and sensible heat fluxes. The flux coefficients are given by

$$\begin{aligned} E_o &= \rho L k_{\text{turb}} (1-r) q_s(T), & a &= L k_{\text{turb}} r \frac{d\rho q_s}{dT}, \\ F_o &= \sigma g(T) T^4, & b &= 4\sigma T^3 + \rho k_{\text{turb}} c_p \end{aligned} \quad (5)$$

where k_{turb} is the turbulent exchange coefficient, σ is the Stefan–Boltzmann constant, and r is the boundary-layer relative

humidity. Much dynamics is hidden in the determination of r , which in the following will be simply taken as a specified parameter. $g(T)$ is a coefficient describing the net infrared radiative cooling of the surface, taking into account the back radiation from the atmosphere into the surface. Above 310 K, g is essentially zero regardless of CO₂, owing to dominance of water vapour. By solving equation (4) for δT_s , one finds the following expression for the evaporation, which in steady state must equal the precipitation

$$E(T) = E_o + (S - E_o - F_o) \frac{a}{a+b} \quad (6)$$

At cold temperatures, both terms become exponentially small owing to the paucity of water in the atmosphere. As temperature increases, E_o increases exponentially and F_o vanishes, but the coefficient $a/(a+b)$ becomes close to unity, limiting the growth of

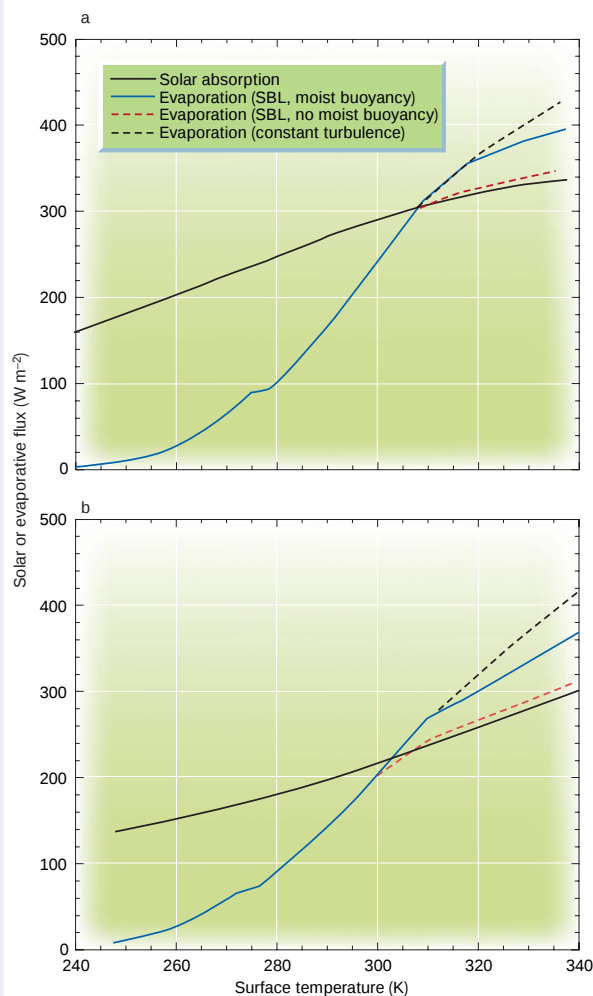


Figure 4 Relation between precipitation and temperature. The mean evaporation from the surface is shown as a function of the surface temperature, together with the amount of solar radiation needed to maintain that temperature. **a**, Low-CO₂ case (100 p.p.m. CO₂ in atmosphere); **b**, high-CO₂ case (0.2 bar). In a steady state, precipitation is equal to evaporation, and 1 W m⁻² of evaporation from a liquid surface corresponds to 1.26 cm yr⁻¹ of liquid-equivalent precipitation (1.11 cm yr⁻¹ for the case of sublimation from an ice surface). The estimates are based on a radiative–convective model, as described in the text. Evaporation is computed for three alternate boundary-layer assumptions: fixed turbulence, stable boundary-layer physics (SBL) including buoyancy effects of water vapour, and stable boundary-layer physics neglecting buoyancy effects of water vapour.

evaporation. However, the exponential increase of E_0 allows evaporation to exceed the available solar radiation in very warm conditions. This is possible in energetic terms because the surface becomes colder than the atmosphere, so that the missing energy can be supplied by sensible heat flux from the atmosphere into the surface. The use of a constant k_{turb} is implausible in this case, as turbulent transfer through the stable boundary layer would be greatly inhibited³⁴. Furthermore, convection would be choked off, limiting the import of dry air that keeps $r < 1$ and sustains evaporation.

It is common practice in climate models to suppress boundary-layer turbulence when the boundary layer becomes stably stratified³⁵. In determining the stability of the boundary layer, there is another factor that needs to be taken into account — water vapour is lighter than dry air. Moisture ‘dissolving’ in dry air can cause it to rise just as salt dissolving in the surface of a tank of fresh water can cause water to sink. This effect becomes particularly significant at warm temperatures, when water vapour makes up a considerable fraction of the mass of saturated air. The buoyancy of water vapour is represented conventionally by virtual temperature, which is the temperature a dry-air parcel would need to have in order to yield the same density (at fixed pressure) as the moisture-laden parcel³⁶. Virtual temperature always exceeds the actual temperature, allowing the boundary layer to be unstable even if the surface is somewhat colder than the overlying drier air. For example, if the boundary-layer relative humidity is 80% while the surface layer is saturated, then surface air at 320 K is positively buoyant when $\delta T_s > -1.44$ K.

In the results presented below, an idealized boundary-layer parameterization is used, which gradually shuts off the turbulent transfer when the virtual temperature of the saturated air in contact with the surface becomes colder than that of the overlying air. For comparison, results neglecting the effect of water vapour on buoyancy, and results with constant k_{turb} , are also shown. Calculations were carried out with a maximum k_{turb} of 0.016 m s^{-1} (corresponding to a wind speed of 8 m s^{-1} in neutrally stratified conditions), and with boundary-layer relative humidity $r = 0.8$.

To complete the problem, the top-of-atmosphere energy budget must also be satisfied. This is accomplished using a radiative–convective model similar to that of the previous section, except that a small (and largely inconsequential) allowance for the effect of δT on OLR is included, and the atmospheric solar absorption is neglected so as to make interpretation of the results more straightforward. Results for a low- CO_2 (100 p.p.m.) and a high- CO_2 (0.2 bar) case are shown in Fig. 4. Because we are interested specifically in understanding the precipitation–temperature relation, the surface temperature is taken as the independent variable and plots show the solar forcing S needed to maintain the stated temperature, along with the associated evaporation. As solar forcing balances OLR in a steady state, the curves of S versus temperature are nothing more than OLR curves of the sort shown in Box 1.

An interesting point is that the solar absorption affects precipitation even for fixed temperature. This is because evaporation cools the surface, and if there is not enough solar energy to offset the cooling, the surface cools sufficiently to limit further evaporation. Consider two planets, each with a surface temperature of 300 K. The first has high CO_2 concentration, and the surface temperature is sustained by a faint sun yielding 218 W m^{-2} of absorbed solar radiation (Fig. 4b). The precipitation in this case is 258 cm yr^{-1} . The second planet has low CO_2 concentration and needs a brighter sun to sustain it, with absorbed solar radiation 291 W m^{-2} (Fig. 4a). This planet is bathed by 307 cm yr^{-1} of precipitation. Neither of these estimates is dependent on stable boundary-layer physics, as the evaporation is not great enough to drive δT_s negative. The planet with the fainter sun has weaker weathering, and so would have to warm to a higher temperature for the weathering to balance the same amount of CO_2 outgassing as would be in equilibrium in the bright-sun planet. This could be part of the reason that snowball states were not ubiquitous early in Earth’s history. When the Sun is faint, the whole hydrologic

cycle is sluggish, and so with less weathering it is harder to draw CO_2 down to levels sufficiently low to initiate the snowball state.

A distinct change in regime occurs when the evaporation exceeds the absorbed solar radiation (Fig. 4), by which point the surface has become colder than the overlying air. Past this point (about 302 K in the high- CO_2 case), precipitation increases more slowly with temperature. Without the destabilizing effects of water vapour buoyancy, evaporation is limited to a value only slightly in excess of the absorbed solar radiation. In this case the silicate–carbonate weathering thermostat breaks down, as increases in temperature fail to increase precipitation (for a fixed solar brightness) and so cannot halt accumulation of CO_2 in the atmosphere. Incorporating the virtual temperature effect helps sustain boundary-layer turbulence and allows precipitation to exceed solar radiation by a greater margin. But the effectiveness of the thermostat is compromised. For example, for a planet in equilibrium with 300 W m^{-2} of solar absorption, increasing CO_2 from 100 p.p.m. to 0.2 bar increases temperature from 307 K to 340 K, but evaporation is increased by only 66 W m^{-2} , yielding only a 22% increase in weathering rate. Neglecting boundary-layer stabilization altogether and keeping turbulence fixed would allow a somewhat greater increase.

A definitive treatment of precipitation in very warm climates would require a much more sophisticated boundary-layer calculation than that presented here. The point of the calculation I have discussed is that the continued operation of the weathering thermostat in very warm climates depends intimately on what is going on in the boundary layer, and on the myriad processes that affect the surface energy budget. These considerations are crucial to the rate of recovery of CO_2 in very warm climates in the aftermath of a snowball Earth deglaciation. Calibration of the precipitation–temperature relation with GCM experiments using fixed surface temperature will yield incorrect results in warm climates, as such calculations do not respect the surface energy budget.

At cold temperatures, in contrast, the surface energy budget is not a significant constraint, and precipitation decreases exponentially in accordance with the decreasing amount of water vapour available in the atmosphere. The reduction of weathering in cold climates is unambiguous, and the weathering thermostat should work well to inhibit runaway glaciation. Initiation of a snowball state proceeds from a temporary breakdown of the thermostat on the cold side. It has been conjectured that anomalous weathering rates could result from a preponderance of tropical continents¹, or from methane-induced precipitation enhancement (D. P. Schrag, R. A. Berner, P. F. Hoffman & G. P. Halverson, personal communication, 2002).

Once a hard snowball is achieved, the tropical temperatures drop to 250 K or less, at which point precipitation rates of only a few centimetres per year are indicated (Fig. 4; GCMs that achieve a hard snowball often yield even less precipitation). Glacial dynamics is driven by a balance between ice flow and precipitation, so it is far from clear that much glacial erosion could be supported by such a small precipitation rate, which is comparable to that found in the very interior of Antarctica today. Geological evidence is unambiguous in indicating active glacial erosion in conjunction with snowball events^{37,38}, but it is possible (although not certain) that the active glaciers are confined to the beginning and end of the glaciated period². At the end of the glaciated period, when enough CO_2 has accumulated to warm the tropics to around 270 K, the precipitation rises to 50 cm yr^{-1} (Fig. 4), and active glaciers are not problematic. Active glaciers at the beginning of the snowball period could arise at times before the tropics are completely frozen.

Weak though it is, the remnant hydrologic cycle during the hard snowball has potentially important consequences for the erosion of ice cover. Because ice thickness is determined ultimately by a small geothermal heat flux, weak evaporation can have a major impact on ice thickness (see ref. 39, although more recent work with improved models of solar absorption in ice indicates substantial modifications to the exploratory results reported there²⁸). Just as a small amount of

evaporation can thin ice, small amounts of precipitation can thicken it, and perhaps more important, increase its albedo. The full implications of the sluggish-snowball hydrologic cycle are only beginning to be explored.

Other planets, other hydrologies

Hydrologic systems analogous to Earth's water-based hydrology can be built upon other condensable greenhouse gases, any of which have multiple roles as greenhouse gas, cloud-forming substance and latent heat transporter. Examples include methane on Titan⁴⁰ and CO₂ on early Mars⁴¹. A unique feature of the CO₂-water system is the interaction between the hydrologic cycle of one greenhouse gas (water) and the weathering chemistry determining the concentration of the other greenhouse gas (CO₂). One wonders whether these are the only substances on which a thermostatic feedback loop can be built, or if other planets in the Universe may have found alternate solutions to the problem of maintaining habitability. □

doi:10.1038/nature01088

- Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A Neoproterozoic snowball earth. *Science* **281**, 1342–1346 (1998).
- Hoffman, P. F. & Schrag, D. P. The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* (in the press).
- Held, I. M. & Soden, B. J. Water vapor feedback and global warming. *Annu. Rev. Energ. Environ.* **25**, 441–475 (2000).
- Pierrehumbert, R. T. in *Mechanisms of Global Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) Geophys. Monogr. Ser. 112 (American Geophysical Union, Washington DC, 1999).
- Dahl-Johnsen, D. et al. Past temperatures directly from the Greenland Ice Sheet. *Science* **282**, 268–271 (1998).
- Lea, D. W., Pak, D. K. & Spero, H. J. Climate impact of late quaternary Pacific equatorial sea surface temperature variations. *Science* **289**, 1719–1724 (2000).
- Lear, C. H., Elderfield, H. & Wilson, P. A. Cenozoic deep-sea temperatures and global ice volumes from Mg/Ca in benthic foraminiferal calcite. *Science* **287**, 269–272 (2000).
- Huber, B. T. Tropical paradise at the Cretaceous poles? *Science* **282**, 2199–2200 (1998).
- Pearson, P. N. et al. Warm tropical sea surface temperatures in the Late Cretaceous and Eocene epochs. *Nature* **413**, 481–487 (2001).
- Trenberth, K. E. & Caron, J. M. Estimates of meridional atmosphere and ocean heat transports. *J. Clim.* **14**, 3433–3443 (2001).
- Manabe, S. & Broccoli, A. J. The influence of continental ice sheets on the climate of an ice age. *J. Geophys. Res.* **90**(C2), 2167–2190 (1985).
- Broccoli, A. J. Tropical cooling at the last glacial maximum: an atmosphere-mixed layer ocean model simulation. *J. Clim.* **13**, 951–976 (2000).
- Hall, N. M. J., Dong, B. & Valdes, P. J. Atmospheric equilibrium, instability and energy transport at the last glacial maximum. *Clim. Dynam.* **12**, 497–511 (1996).
- Sokolov, A. P. & Stone, P. H. A flexible climate model for use in integrated assessments. *Clim. Dynam.* **14**, 291–303 (1998).
- Jenkins, G. S. & Smith, S. R. GCM simulations of Snowball Earth conditions during the late Proterozoic. *Geophys. Res. Lett.* **26**, 2263–2266 (1999).
- Poulsen, C., Pierrehumbert, R. T. & Jacob, R. Impact of ocean dynamics on the simulation of the Neoproterozoic “Snowball Earth”. *Geophys. Res. Lett.* **28**, 1575–1578 (2001).
- Barron, E. J., Fawcett, P. J., Peterson, W. H., Pollard, D. & Thompson, S. L. A simulation of mid-Cretaceous climate. *Paleoceanography* **10**, 953–962 (1995).
- Bush, A. B. & Philander, S. G. H. The late Cretaceous: simulation with a coupled atmosphere-ocean

- general circulation model. *Paleoceanography* **12**, 495–516 (1997).
- Huber, M. & Sloan, L. C. Heat transport, deep waters, and thermal gradients: coupled simulation of an Eocene Greenhouse Climate. *Geophys. Res. Lett.* **28**, 3481–3484 (2001).
- Huber, M. & Sloan, L. C. Climatic responses to tropical sea surface temperature changes on a “greenhouse” Earth. *Paleoceanography* **15**, 443–450 (2000).
- Kirk-Davidoff, D. B., Schrag, D. P. & Anderson, J. G. On the feedback of stratospheric clouds on polar climate. *Geophys. Res. Lett.* (in press).
- Hyde, W. T., Crowley, T. J., Baum, S. K. & Peltier, W. R. Neoproterozoic ‘snowball Earth’ simulations with a coupled climate-ice sheet model. *Nature* **405**, 425–429 (2000).
- Chandler, M. A. & Sohl, L. E. Climate forcings and the initiation of low-latitude ice sheets during the Neoproterozoic Varanger glacial interval. *J. Geophys. Res.* **105**, 20737–20756 (2000).
- Caldeira, K. & Kasting, J. F. Susceptibility of the early Earth to irreversible glaciation caused by carbon-dioxide clouds. *Nature* **359**, 226–228 (1992).
- Kasting, J. F., Pollack, J. B. & Ackerman, T. P. Response of Earth's atmosphere to increases in solar flux and implications for loss of water from Venus. *Icarus* **57**, 335–355 (1984).
- Pierrehumbert, R. T. & Erick, C. On the scattering greenhouse effect of CO₂ ice clouds. *J. Atmos. Sci.* **55**, 1897–1903 (1998).
- Briegleb, B. P. Delta-Eddington approximation for solar radiation in the NCAR Community Climate Model. *J. Geophys. Res.* **97**, 7603–7612 (1992).
- Warren, S. G., Brandt, R. E., Grenfell, T. C. & McKay, C. P. Snowball Earth: ice thickness on the tropical ocean. *J. Geophys. Res.* (in the press).
- Hay, W. W., DeConto, R. M. & Wold, C. N. Climate: is the past the key to the future? *Geol. Rundsch.* **86**(2), 471–491 (1997).
- Sagan, C. & Mullen, G. Earth and Mars—evolution of atmospheres and surface temperatures. *Science* **177**, 52–56 (1972).
- Walker, J. C. G., Hays, P. B. & Kasting, J. F. A negative feedback mechanism for the long-term stabilization of Earth's surface-temperature. *J. Geophys. Res.* **86**(Nc10), 9776–9782 (1981).
- Berner, R. A. GEOCARB II: a revised model of atmospheric CO₂ over Phanerozoic time. *Am. J. Sci.* **294**, 56–91 (1994).
- Berner, R. A. & Kothavala, Z. GEOCARB III: a revised model of atmospheric CO₂ over phanerozoic time. *Am. J. Sci.* **301**, 182–204 (2001).
- Vogelezang, D. H. P. & Holslag, A. A. M. Evaluation and model impacts of alternative boundary-layer height formulations. *Bound. Layer Meteorol.* **81**, 245–269 (1996).
- Kiehl, J. T. et al. The National Center for Atmospheric Research Community Climate Model: CCM3. *J. Climate* **11**, 1131–1149 (1998).
- Emanuel, K. A. *Atmospheric Convection* (Oxford Univ. Press, New York, 1994).
- Christie-Blick, N., Sohl, L. E. & Kennedy, M. J. Considering a Neoproterozoic snowball Earth. *Science* **284**, 1087 (1999).
- McMechan, M. E. Vreeland diamictites-Neoproterozoic glaciogenic slope deposits, Rocky Mountains, northeast British Columbia. *Can. Bull. Petrol. Geol.* **48**, 246–261 (2000).
- McKay, C. P. Thickness of tropical ice and photosynthesis on a snowball Earth. *Geophys. Res. Lett.* **27**, 2153–2156 (2000).
- Lunine, J. I., Lorenz, R. D. & Hartmann, W. K. Some speculations on Titan's past, present and future. *Planet Space Sci.* **46**, 1099–1107 (1998).
- Forget, F. & Pierrehumbert, R. T. Warming early Mars with carbon dioxide clouds that scatter infrared radiation. *Science* **278**, 1273–1276 (1997).
- Kasting, J. F. Runaway and moist greenhouse atmospheres and the evolution of Earth and Venus. *Icarus* **74**, 472–494 (1988).

Acknowledgements

I am indebted to K. Trenberth, M. Huber and C. Poulsen for providing data used in this review, and for much other valuable assistance. P. Hoffman and D. Schrag introduced me to the snowball Earth problem, and our discussions on this subject have continued over the years; insofar as I understand anything at all about the phenomenon, much credit is due to them. I also benefited from comments by R. Alley, T. Schneider and S. Warren. I had the further advantage of meetings and discussions carried out as part of my participation in the NOAA Panel on Abrupt Climate Change, for which the support of the National Oceanographic and Atmospheric Administration is gratefully acknowledged.

Links between climate and sea levels for the past three million years

Kurt Lambeck^{*,†}, Tezer M. Esat^{*} & Emma-Kate Potter^{*}

^{*}Research School of Earth Sciences, Australian National University, Canberra 0200, Australia

[†]Swedish Research Council and Lund University, Tornavägen 13, S223 63, Sweden

The oscillations between glacial and interglacial climate conditions over the past three million years have been characterized by a transfer of immense amounts of water between two of its largest reservoirs on Earth — the ice sheets and the oceans. Since the latest of these oscillations, the Last Glacial Maximum (between about 30,000 and 19,000 years ago), ~50 million cubic kilometres of ice has melted from the land-based ice sheets, raising global sea level by ~130 metres. Such rapid changes in sea level are part of a complex pattern of interactions between the atmosphere, oceans, ice sheets and solid earth, all of which have different response timescales. The trigger for the sea-level fluctuations most probably lies with changes in insolation, caused by astronomical forcing, but internal feedback cycles complicate the simple model of causes and effects.

The onset of Pleistocene glaciations occurred about three million years ago with formation of permanent ice sheets at high northern latitudes. As the ice sheets waxed and waned, the concomitant fall and rise of sea level left direct evidence for the intensity and timing of glacial cycles. The major sea-level cycles occur at intervals of ~100,000 years (100 kyr) over the past ~800 kyr, with maximum amplitudes of 120–140 m, involving changes in ice volume of 50–60 million km³. Superimposed on these are lesser cycles of a few tens of thousands of years and shorter duration.

Direct, accurate and high-resolution observations of past sea levels exist only for the last glacial cycle, from about 130 kyr ago to the present. Knowledge of ice volume for the earlier periods rests increasingly on proxy indicators contained within oozes of calcareous skeletons of marine foraminifera shells. These record the isotopic and chemical signatures of the water in which they lived, and the ratio of two oxygen isotopes ¹⁸O/¹⁶O is particularly important. It is the lighter isotope that is preferentially evaporated, whereas during precipitation the heavier isotope condenses preferentially. Thus, as water vapour is transported poleward this

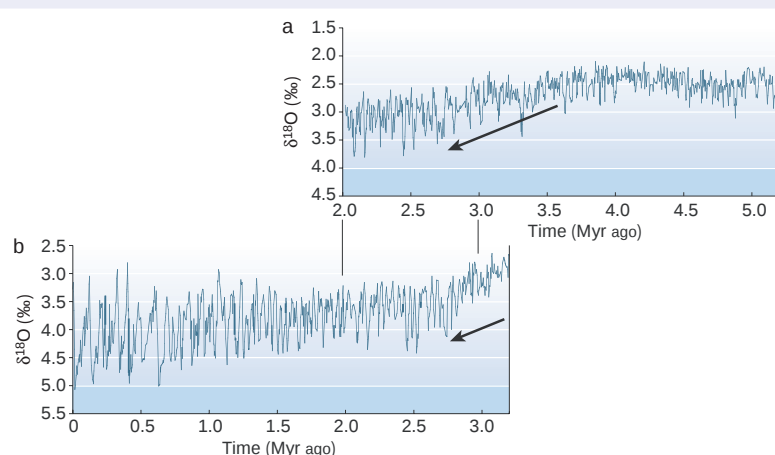
ratio shifts to lesser values and is reflected in the isotopic signature of high-latitude snow. At times of major glaciations, the oceans become depleted in the lighter oxygen isotope.

The isotope ratio $\delta^{18}\text{O} = \{(^{18}\text{O}/^{16}\text{O})_{\text{sample}} / (^{18}\text{O}/^{16}\text{O})_{\text{standard}} - 1\}$, expressed as parts per thousand, is therefore believed to be an indicator of global ice volume — low values indicate small ice volumes and hence globally warm conditions, and high values imply large ice sheets and low temperatures^{1,2}. In foraminifera, this ratio is also a function of local seawater temperature and independent observations become important for establishing not only the links between the sea-level fluctuations and climate, but also for calibrating the proxy indicators. Accumulating evidence reveals a complex pattern of change throughout Pleistocene time, particularly for the last glacial cycle, with a variable role for astronomical forcing and an interaction between oceans, atmosphere, ice and solid earth that is rich in its implications for climate change.

Onset of glacial cycles in sea level

Records of foraminifera have been preserved in open-ocean marine sediments from two localities over the past 5 million years (Myr)^{3,4} (Fig. 1). The record before ~3 Myr is

Figure 1 $\delta^{18}\text{O}$ records for the past 5 Myr from two sediment cores. **a**, Site 999 in the Caribbean Sea⁴; **b**, Site ODP-607 from the equatorial Atlantic³. For the early part of the record (before ~3.2 Myr ago) the average $\delta^{18}\text{O}$ is low and the oscillations are of relatively small amplitude. This is indicative of globally warm conditions without major ice sheets in the Northern Hemisphere. At ~2.7 Myr ago both the oscillations and the mean value are larger, with a principal periodicity of ~40 kyr. After ~0.8 Myr ago the major oscillation occurs with a periodicity of ~100 kyr.



characterized by low $\delta^{18}\text{O}$ and the early period is one of relatively high global temperatures with little ice locked up in polar caps. After ~ 3 Myr the oscillations become larger and indicate onset of glaciation in the Northern Hemisphere, with successive glaciations becoming progressively more intense. This onset coincides with a marked increase in lithic material in high-latitude North Atlantic sediment cores^{4,5} carried by the first calving of Northern Hemisphere ice into the Atlantic. At ~ 0.8 Myr ago the pattern of $\delta^{18}\text{O}$ oscillations shifts from the principal periodicity of 40 kyr to 100 kyr (ref. 6). The 100-kyr pattern becomes established in the sea-level record during this interval^{7–10}.

Growth of large ice sheets requires warm temperatures during winters to enhance moisture transport to high latitudes, and cool temperatures during summers to prevent melting of snow. Surface water transports heat efficiently and a northward penetration of warm surface currents can encourage ice-sheet development, while suppression of these currents discourages ice growth. Thus changes in climate, ocean circulation and ice-sheet growth and decay are closely linked^{11–14}. A number of factors can modify this circulation. Changes in tectonic configuration of the continents can drive warm surface currents towards or away from high latitudes. The closure of the Isthmus of Panama may have been important in this context, because as the waterway between the two Americas narrowed, warm surface Atlantic currents were probably deflected northwards. The timing of the final stages of closure between about 3.6 and 2.7 Myr ago^{4,15} is sufficiently close to the onset of Northern Hemispheric glaciation to make an attractive hypothesis: that the tectonically driven change in the Atlantic circulation leads to increased moisture transport to high latitudes and to the development of the northern ice sheets. Once established, the ice sheets can increase the supply of cold water to the oceans, which feeds back into the surface currents. Superimposed upon this tectonic–ocean–ice interaction will be any external forcing of the climate.

Under the influence of gravitational interactions between the Earth, Sun and Moon, the insolation at the upper atmosphere varies cyclically¹⁶ and the same periodicities are observed in the $\delta^{18}\text{O}$ records¹⁷. Known as the Milankovitch forcing, this indicates that astronomical forces exert a strong ordering influence on climate^{16,18,19}, although some significant differences do exist. For example, the climate periodicity of 100 kyr is only weakly present in orbital frequencies, and a strong 413-kyr cycle seen in insolation predictions is not reflected in climate variability²⁰. On a timescale of a few million years the orbital and rotational motions in the Solar System are stable²¹, and if astronomical forcing is the only agent for change then the same climate behaviour is expected throughout, in contrast to the fundamental changes seen at ~ 3 and ~ 0.8 Myr ago. Thus, processes and feedbacks other than astronomical motions are likely to be important in influencing climate variability^{11,13,20}. Suggestions for feedback modes are many, but satisfactory solutions are few. Where there is agreement is that changes in ocean circulation, whose response to external forcing contains both short (~ 1 kyr) and longer (3–5 kyr) time constants, can modify climate response to external forcing and feedback mechanisms^{22,23}.

Numerous explanations have been advanced for the dominant 100-kyr cycle in the climate record. Most involve the long response times associated with massive ice sheets^{24,25}, others involve forcing by orbital inclination²⁶. However, atmospheric variables extracted from the Vostok ice core and from marine cores follow a 100-kyr cycle in phase with astronomical forcing, indicating that this cycle is not related to ice-sheet dynamics²⁷. Instead, it seems to be a response (with appropriate time lag) to forcing by air temperature. This leaves unanswered the cause of the temperature forcing and why it started only ~ 0.8 Myr ago. But, it means that the relationship between $\delta^{18}\text{O}$ and sea-level fluctuations is more complex than sometimes assumed, reinforcing the need to establish an independent sea-level record free of proxy indicators. Availability of increasingly accurate observations of past sea-level positions should lead to answers to

some of the questions raised by the proxy records and hence to an improved understanding of the processes that have driven climate change in the past.

Earlier glacial cycles and sea level

Observational records of sea level consist of positions and ages of remnants of former shorelines. Coral reefs^{9,10,28} and speleothems (reprecipitated carbonates)²⁹ provide a chronology of mainly interglacial periods and these confirm the recurrence of the ~ 100 -kyr cycle. The amplitudes of sea-level change during the early interglacials are poorly known because many of the locations where records are preserved are also subject to tectonic uplift. Thus it remains impossible to say with certainty whether there was more or less ice globally than today during any of the past interglacials and whether the behaviour of the Greenland and West Antarctic ice sheets differs from interglacial to interglacial. The observed timing of earlier interglacials, at ~ 630 kyr ago¹⁰, ~ 330 kyr ago^{10,30} and ~ 200 kyr ago²⁹, is consistent with the hypothesis of astronomically forcing. But, other than for the last interglacial, it has not yet been possible to define with precision the timing of the onset and termination of these periods or to establish in detail the phase relationship(s) between Milankovitch forcing, sea level and oxygen isotope signals for the same interval. Current evidence suggests that the last two interglacials were of approximately equal length (~ 12 kyr)^{29,31}.

Sedimentological evidence has provided some estimate of sea levels during the major glacials preceding the interglacials^{32–34}, and global ice volumes during successive glacial maxima seem to have been of similar magnitude to that of the Last Glacial Maximum (LGM; see below). One possible exception is at ~ 450 kyr ago, when ice volumes may have been some 15% greater than during the last glaciation³², and greater than at any other time in the past 3 Myr. Determining where this ice was will remain uncertain until a coherent chronology for the terrestrial record of the early glaciations is established.

The last glacial cycle

In areas of tectonic uplift, sea-level markers have occasionally been elevated beyond the reach of the destructive forces of succeeding cycles. However, because the quality of the exposed record often deteriorates as a result of weathering, a quasi-continuous sea-level record is still only available for the last glacial cycle, from about 130 kyr ago to the present. This reveals a much finer structure than do the older fragments, one that is diagnostic of higher-frequency interactions between climate, ocean and ice.

Coral reefs form particularly good records. Coral growth proliferates when the rate of sea-level rise equals or exceeds the rate of land uplift, but when sea-level rise cannot keep up only patchy and thin reef veneers develop. Important examples are from Barbados³⁵, Sumba⁹ and Papua New Guinea^{8,36,37}. Uplift at Huon Peninsula, Papua New Guinea, is as high as 4 mm yr^{-1} , resulting in a more detailed record than is available from other localities. Huon terraces of last interglacial age are now at elevations up to 400 m, whereas in tectonically stable areas such reefs occur near modern sea level³¹. Comparison between the two sites yields estimates for the rate of tectonic uplift, while evidence from the latter sites is used to establish the timing of the interval when ice volume was last near its present volume^{28,31}. Information from river delta sequences at Huon constrains the sea-level position of intervening lowstands³⁴. Uranium-series dating of the corals^{38,39} completes the process, leading to a relation between age, sea level and height⁴⁰ (Fig. 2). Although the Huon data have already received considerable attention, new high-resolution results are providing fresh insights into the behaviour of ice sheets, thermohaline circulation and sea level^{41,42}.

The last glacial cycle is divided into marine oxygen isotope stages (MIS) 1–5 based on distinct patterns of fluctuation in $\delta^{18}\text{O}$ in marine sediments⁴³. MIS-5 includes the last interglacial (substage MIS-5e) and a series of oscillations between increasingly cold stadials and relatively warm interstadials (substages 5d–5a; Fig. 2). MIS-4 and -3

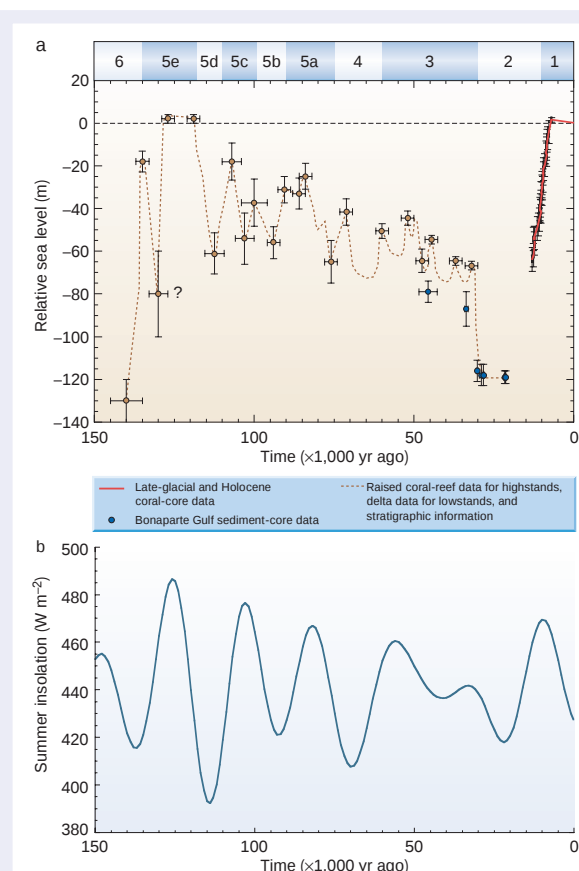


Figure 2 Relative sea level and insolation for the last glacial cycle. **a**, The relative sea-level curve for the last glacial cycle for Huon Peninsula⁴⁰ supplemented with observations from Bonaparte Gulf, Australia⁷³. Error bars define the upper and lower limits. The timescale is based on uranium-series ages of corals older than about 30 kyr (refs 36–38) and on calibrated radiocarbon ages of the younger corals and sediments (all U-Th ages here and in Fig. 3 have been subjected to strict evaluation criteria²¹). The main oxygen isotope stages MIS-5 to MIS-1, including the substages of MIS-5, are identified. The results represent conservative estimates, with each oscillation identified being supported by several precisely dated points from one or more reef section, by the stratigraphic relationship of the sampled corals to the reef crest position, and by models of reef growth in an environment of changing sea levels⁴². The local minima in sea level during MIS-5 are based on the positions of stream deltas relative to the reefs defining local highstands³⁴. These lowstands have not been dated, but their positions are consistent with modelling results of reef growth⁴². Although it is not normally permissible to combine sea-level data from different localities without first correcting for differential isostatic effects, in this particular case the Huon and Bonaparte locations are at continental margins far from former ice loads and the differences in isostatic corrections are less than the uncertainties in the data⁷⁵. Thus the Bonaparte data are assumed to be valid for Huon within the error bars indicated. **b**, Insolation curve for July at 65° N.

correspond to the period leading into the maximum glaciation and represents a time of stadials, of increasing intensity, punctuated by brief interstadials. MIS-2 includes the LGM followed by a period of rapid ice decay. The past ~12,000 years defines MIS-1 or the Holocene, when ice volumes and climate were largely similar to those of today.

Into and out of the last interglacial period

The end of the last interglacial (MIS-5e) marks a return to glacial conditions at high northern latitudes and, along with substages 5d to 5a, forms a possible analogue for the termination of the Holocene.

Shorelines and coral reefs formed during MIS-5e are a few metres higher than modern sea level and the highest levels may have occurred midway through the period³¹. These higher values cannot be related directly to differences in ice volume between this period and the Holocene because of the isostatic adjustment of the Earth in response to ice-water loads before and after reef formation^{44,45}. What can be concluded is that it is unlikely that there were major differences in ice volume (<2–4 m in equivalent sea level) between the two interglacials.

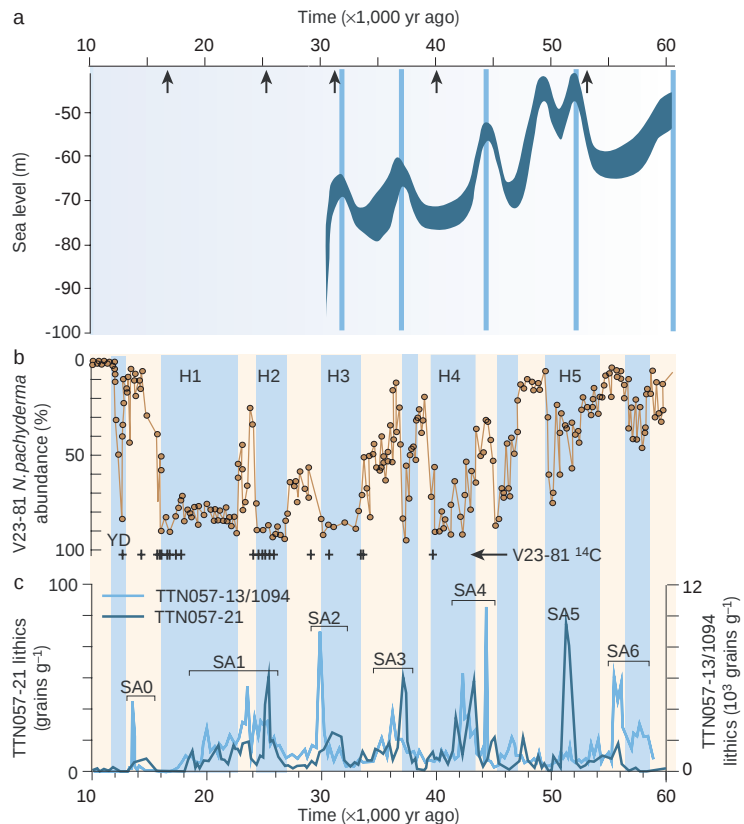
Sea levels during the last interglacial reached present values for the first time around 130 kyr ago^{31,46}. However, if the analogy with the Holocene is valid, interglacial temperatures in mid-latitudes will have been attained as much as 6,000 years before the sea approached its modern level, and this is indeed supported by terrestrial $\delta^{18}\text{O}$ records⁴⁷ and other marine deposits⁴⁸. The period leading out of the penultimate glacial maximum and into the last interglacial remains poorly constrained. Sea levels during this maximum were about the same as during the LGM^{33,34}, but the timing is not established quantitatively. The subsequent sea-level rise seems to have been rapid, possibly with a swift, large-amplitude oscillation before modern levels were reached^{37,46}.

At Huon and Barbados, well-developed monolithic reef structures formed during the major interstadials MIS-5a and MIS-5c^{8,35} when global sea levels were 20–30 m lower than today. Relative highstands corresponding to MIS-5a have also been recorded in Florida^{49,50} and Bermuda⁴⁹, but their elevations are affected by isostatic adjustment to past ice loads more than at Barbados and Huon. After differential isostatic corrections are applied, the results are consistent within the uncertainties of the observations and the model corrections⁴⁵. During the stadials MIS-5b and MIS-5d, sea levels have been as much as 40–60 m below present levels⁴⁰. For comparison, the Scandinavian and Laurentian ice sheets at maximum limits contained enough ice to lower sea level by ~12–15 m and ~60–80 m, respectively⁵¹. Thus, substantial ice sheets existed during the MIS-5 stadials.

Few pre-LGM ice margins and shorelines have been preserved that provide even approximate constraints on thickness or location of earlier glaciations and, where they have been identified, there is debate about their age. In Scandinavia, MIS-5d is recognized as a cool period during which ice was restricted to the high ground of Norway and Sweden^{52,53}, contributing only a few metres to the change in global sea level. The glaciation during stage MIS-5b was more extensive, but still only a small fraction of its subsequent maximum volume. Only by MIS-4 did a large ice sheet form over Scandinavia. In between these stadials, the ice retreated to the mountains or disappeared entirely⁵⁴. Instead, the major Eurasian glaciation began over Arctic Russia with a large ice sheet over the Kara Sea and western Siberia⁵⁵. Field estimates of ice margins during stages 5b and 4, including observations of ice overriding the higher mountains of the region⁵⁵, indicate a glaciation comparable in size to the LGM Scandinavian ice. When ice grew over Scandinavia during MIS-4, to the east it started to retreat and by the coldest time (MIS-2) the Kara Sea shelves and Arctic Russia were largely free of glacier ice⁵⁵. Likewise, in North America the early MIS-5 glaciations initiated in Arctic Canada and successive glaciations advanced progressively southwards during MIS-5d to -4 while decaying in the north⁵⁶.

The mechanisms by which the Eurasian and North American ice sheets developed following MIS-5e remain uncertain and speculative. However, the early growth of ice sheets at high latitudes indicates an enhanced moisture supply to the Arctic regions compared to the present, possibly because during warmer parts of the last interglacial, permanent Arctic sea-ice cover was less extensive than during subsequent periods. This introduces moisture into the polar areas and enables snow to accumulate without requiring a significant lowering of temperature. As falling global temperatures initiate ice build-up in Scandinavia, the sea ice returns and a precipitation shadow develops that starves the eastern region of further snow. A similar scenario could be envisaged for North America. Open sea conditions at high latitudes during the late stages of the last interglacial allow ice to

Figure 3 Comparison of MIS-3 sea-level oscillations and Atlantic sediment records. **a**, Sea-level curve for MIS-3 coral terraces from Huon Peninsula, Papua New Guinea. Upper and lower bounds are shown. **b**, *Neogloboquadrina pachyderma* (s.) abundances from the North Atlantic core V23-81⁵⁹ plotted on the same calendar timescale. The original radiocarbon ages^{57,59} from the depths in the core marked by (+) have been converted to calendar ages. The extent of major Heinrich events identified in the North Atlantic cores, corresponding to cold periods, are within the colour bands. The arrows identify the mean timing of Heinrich events⁶². **c**, IRD from two South Atlantic cores TTN057-21 and TTN057-13/1094 (ref. 60). Within the uncertainties of the calendar age determinations, the South Atlantic IRD peaks correspond to the cold events in the North Atlantic.



develop initially over Arctic Canada. The subsequent cooling, and the southern march of the North American ice sheet, leads to a precipitation shadow causing the northern ice to decay at the same time as the maximum southern limit is reached. Thus it is tempting to suggest that onset of a glacial cycle after a prolonged interglacial requires the break up of Arctic sea ice and enhanced exposure of Arctic air to relatively warm surface water. The accumulation of snow and ice increases albedo at high latitudes, which then feeds back into the global climate system. A test of this hypothesis would be to demonstrate that temperature during the MIS-5 fluctuations lags the fall in sea-level.

Leading into the maximum glaciation

Sea level after MIS-5 did not fall uniformly towards its LGM level, but oscillated rapidly with amplitudes of 10–15 m about every 4,000 years (the approximate duration of a half cycle^{36,41}; Fig. 2). There is a clear visual distinction between the MIS-3 and MIS-5 reefs at Huon, the former being less developed, more numerous and possessing more complex sub-structures, a distinction that is mirrored in the $\delta^{18}\text{O}$ records. The implication is that fluctuations in climate and its ice-volume response were more rapid during MIS-3 than during MIS-5.

Five highstands have been identified within the MIS-3 interval, centred on 32, 36, 44, 49–52 and 60 kyr ago (Fig. 3). Periods of high abundance of the planktonic foraminiferan *Neogloboquadrina pachyderma* sinistral (left-coiling) in the North Atlantic core V23-81 closely match the $\delta^{18}\text{O}$ variations and correspond to cold water temperatures⁵⁷. These periods also coincide with episodes of ice-rafted debris (IRD) deposits in both the North^{58,59} and South Atlantic (for example, peaks SA0 to SA6)⁶⁰ (Fig. 3). Variations in IRD abundances in the two hemispheres are in agreement except for the additional peak at about 37 kyr ago in the South Atlantic record. There is, however, a sharp enhancement in *N. pachyderma* abundance in core

V23-81, indicating that this southern IRD event occurs within a cold interval in the Northern Hemisphere. Periods of cooler upper-ocean temperature and increased IRD define Heinrich events (H1 to H6)^{58,61} of which the last four occurred during MIS-3 (ref. 62; and see review in this issue by Rahmstorf, pages 207–214).

Heinrich events are attributed to instabilities in the ice sheets once they have grown to continental dimensions, resulting in iceberg discharge. The increased supply of fresh cold water, from melting icebergs at high-latitudes in the North Atlantic, is believed sufficient to dilute the warm, saline surface waters and prevent them from sinking^{63,64}. The consequential slow-down or interruption in the thermohaline circulation causes a cold snap, reduction in freshwater supply to the ocean and the eventual re-establishment of North Atlantic Deep Water formation⁶⁵. The cycle is completed over the next several thousand years when renewed moisture supply and snow precipitation leads to re-growth of the ice sheets.

The chronology of the sedimentary records after ~50 kyr ago is usually related to the radiocarbon timescale, whereas corals are usually dated using uranium-series methods. Radiocarbon ages become increasingly uncertain with age, not only because of their decreasing levels of ¹⁴C, but also because ¹⁴C production in the atmosphere varies with time. In particular, anomalous rates of ¹⁴C production and large excursions from a uniform timescale occur during Heinrich events^{41,66,67}. Uranium-series ages, although more accurate than radiocarbon ages for this time interval, may also not be as accurate as formal precision estimates suggest because of the possibility of coral diagenesis^{28,31}.

Although accurate phase relations between sediment records and sea-level oscillations cannot yet be established, the following three conclusions can at least be drawn from a comparison of the sediment and coral results⁴¹. First, periods of enhanced reef development recorded by the Huon terraces correlate with cold intervals; second,

reef growth occurs in response to rising sea level caused by ice-sheet collapse during cold periods; and third, ice discharge at times of Heinrich events raises global sea level by 10–15 m. If these conclusions are accepted then the best definition of the timing of Heinrich events is established not by the uncertain ^{14}C chronology, but by the Huon uranium-series reef ages.

The amplitude of sea-level rise represents 15–25% of Laurentide ice volume or the total volume of Scandinavian LGM ice. Thus the Heinrich events are associated with significant and rapid changes in the ice sheets. Different locations have been identified for the source regions of IRD. Some of the thickest and most rapidly deposited layers are close to the Hudson Strait, and the Laurentide ice sheet is a chief source⁶⁸. But deposits of similar timing originate from the European ice sheet⁶⁹ and the correspondence between South and North Atlantic deposits suggests that coeval fluctuations in ice volume occur in Antarctica⁷⁰. Thus there must be strong and swift interactions between the major ice sheets in both hemispheres, in which the collapse of one ice sheet raises sea level sufficiently to destabilize those margins of the others where ice advanced onto the shelves⁷¹. According to this scenario, rapid sea-level rises need not be sought in the collapse of only a single ice sheet.

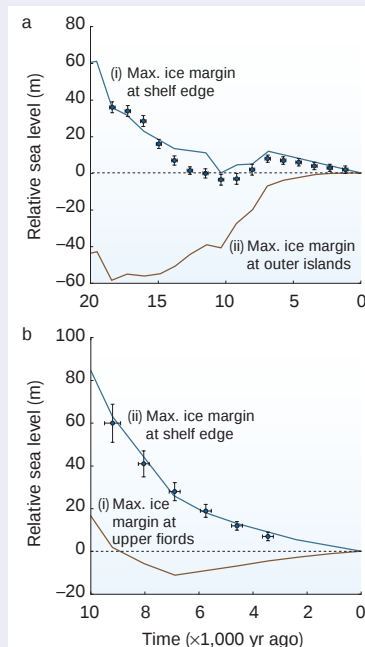
The Last Glacial Maximum

The lowest sea levels at any time during the last glacial cycle occurred from ~30 kyr ago to ~19 kyr ago⁷²; this period constitutes the LGM when land-based ice volumes were $\sim 55 \times 10^6 \text{ km}^3$ greater than present^{73,74}. Onset of the LGM was rapid, with sea level falling ~30–40 m within 1,000–2,000 years⁷⁵. The growth in ice is consistent with evidence from Scandinavia that extensive ice-free conditions ended before ~30 kyr ago⁷⁶ with rapid expansion of ice from Norway and Sweden over the North Sea⁷⁷ and Finland⁷⁸. Similar ice-sheet expansion is required in North America or Antarctica to explain the observed magnitude of sea-level change. The rapidity of sea-level lowering events, including those recorded during MIS-3 Heinrich cycles, indicate that ice sheets can expand quickly under favourable conditions of moisture supply.

Observations of local sea level at and within the former ice margins provide constraints on the contributions of individual ice sheets to the total ice volume^{79,80}. During glacial loading, the crust beneath the ice is depressed and rebounds slowly when ice is removed, at a rate determined by the earth's rheology and by the ice-load history. By exploiting the spatial and temporal variability of the sea-level response, some separation of earth and ice parameters can be achieved. Near the former centre of glaciation, the contribution of the crustal rebound dominates sea-level change and results in a characteristic exponentially decreasing function. For sites near but within the maximum ice margin, the rebound and global-rise signals are of comparable magnitude but opposite in sign, producing the typical ice-margin signal from Vesterålen, Norway (Fig. 4a). Here organic sediments indicate that the offshore islands were partly ice free as early as 21.5 kyr ago when local sea level stood 35 m higher than today⁸¹. If the LGM ice never extended beyond this location, the predicted levels are considerably lower than observed because the rebound is insufficient to overcome the rise from the added global meltwater. Only if the ice stood at the shelf edge, ~50 km offshore and ~1,000–1,500 m thick, do predicted levels begin to match the observed values⁸⁰.

A similar conclusion is drawn from age–height observations of shorelines on Baffin Island, Canada. In some reconstructions of the North American ice sheet, the maximum ice margin does not reach the coast during the LGM; rather the ice limit remains near the heads of the fiords⁸². Sea-level curves are similar to those observed along the Norway coast (Fig. 4b), but if ice reached only the upper fiord regions the predicted values are mostly below sea level. Only if ice extended onto the shelf do predictions agree with observed evidence. A recent re-evaluation of geomorphological evidence, supplemented with cosmogenic age data, concludes that the ice did indeed extend offshore in the south, but not in the north⁸³. Sea-level analyses for

Figure 4 Sea-level fluctuations near former ice margins. **a**, Data from Andøya, Vesterålen, northern Norway. The observed evidence⁸¹ is shown by the circles with error bars. Predicted values⁸⁰ are shown for two ice models: (i) where the maximum ice margin is located at the continental edge at the time of the LGM and before the site became ice free; and (ii) where the maximum ice margin extended only to the present coast. **b**, Data from Baffin Island, Canada. The observed sea levels⁷⁹ are shown by the open circles. The model predictions are for: (i) an ice model in which the ice did not expand beyond the upper fiords of eastern Baffin Island⁸¹; and (ii) a model in which the ice margin has been shifted 80 km eastwards and



onto the shelf, retaining the same height profile as in (i). Only in the latter case is there good agreement between observations and predictions. The expansion of the ice onto the shelf is consistent with the re-interpretation of morphological evidence⁸³ from southern and central Baffin Island. It is not consistent, however, with evidence from northern Baffin Island where the few available sea-level indicators also imply thick LGM ice cover onto the adjacent shelf, but the morphological re-interpretation is that the ice cover was more restricted⁸³. We suspect that further re-interpretation of the field data will be necessary before this discrepancy is resolved.

Svalbard⁸⁴, Greenland⁸⁵ and East Antarctica⁸⁶ lead to similar conclusions that, at the height of glaciation, the ice sheets extended offshore, in most cases as far as the shelf margin where they were susceptible to rapid reductions in volume.

The thickness of the ice can be constrained by the amplitudes of local sea level. The rebound predicted for the steeply domed CLIMAP ice model⁸⁷ is greater than observed for a wide range of plausible mantle rheologies, indicating that the ice thickness is overestimated^{79,80}. But if the volumes of the individual ice sheets are scaled down uniformly such that predicted and observed rebound agrees, then the total ice volumes become inadequate to explain the observed LGM sea levels. However, the mostly late- and post-glacial rebound data do not adequately constrain ice volumes at the LGM and the observations are also consistent with rebound models in which ice sheets that were initially steeply domed experienced early rapid reduction⁸⁸. The observed rapid sea-level rise of about 10–15 m after 19 kyr ago⁷³ coincides with a substantial IRD influx, mainly of Scandinavian origin, into the Irminger Basin⁸⁹, which suggests that a major change occurred in the Scandinavian ice sheet at this time. Alternatively (or additionally), this sea-level rise could have an Antarctic origin, as the $\delta^{18}\text{O}$ signal from the Southern Ocean⁹⁰ and the Antarctic Vostok ice core⁹¹ indicates that warming in the Southern Hemisphere started abruptly at about 20 kyr ago and that a partial dispersal of the grounded shelf ice followed soon after.

The late- and post-glacial period

The post-LGM record from localities far from the ice sheets indicates that sea-level rise was not temporally uniform (Fig. 5). On some shallow continental shelves, such as within the Persian Gulf, now-submerged shorelines developed that are indicative of constant sea level,

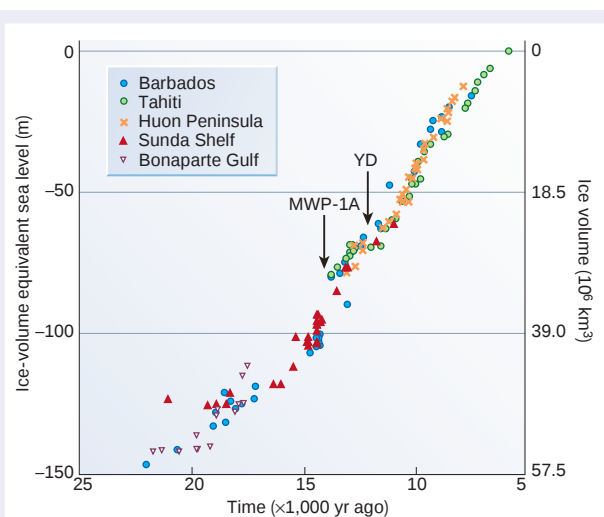


Figure 5 Changes in global ice volume from the time of the LGM to the present. The figure shows ice-volume equivalent sea level for the past 20 kyr based on isostatically adjusted sea-level data from different localities^{73,93–96}. Because of spatial variability of the sea-level response to the glacial and water loading, sea-level observations from different localities should not be combined into a single sea-level curve unless the isostatic effects can be shown to be similar. The ice-volume equivalent sea-level function used here corresponds to the observed sea levels corrected for these effects⁷⁵. It relates to the change in total ice volume, with respect to the present, of continent-based ice and any ice grounded on the shelves. With one exception, the results indicate an ice volume at the LGM that was $\sim 55 \times 10^6 \text{ km}^3$ greater than at present^{74,75}. The error bars, not shown, are typically 0.1–0.15 kyr in calendar years and 5 m or less in position. MWP-1A refers to the timing of the meltwater pulse at ~ 14 kyr ago. At the Younger Dryas (YD) at ~ 12 kyr ago, sea-level rise may have momentarily halted.

particularly about the time of the Younger Dryas⁹². This is also seen in Fig. 5 as a short-duration period at ~ 12 kyr ago when ice volumes remained constant. A period of sustained global melting occurred from ~ 19 –7 kyr ago, but periods of more rapid rise can also be identified within this interval⁹³. A gap in the Barbados record occurs at ~ 14 kyr ago, at a time when sea level rose at a rate of about 40 mm yr^{-1} (ref. 94).

The record is not well defined for this latter period, possibly because of a sampling artefact or because the coral growth could not keep pace with the rapid sea-level rise, but a similar hiatus has been identified in the coral record from Tahiti⁹⁵. Furthermore, data from the Sahul Shelf⁹⁶, which are free from ‘keep-up’ considerations, support a rapid rise of about $\sim 15 \text{ m}$ between 16–14 kyr ago, although data of greater temporal resolution are again desirable. This meltwater pulse, MWP-1A⁹⁴, reflects a rapid decay of ice sheets, but no consensus on the source has been reached, with contributions from North America, Scandinavia and the Barents Sea, and Antarctica having been proposed^{97,98}. Estimates vary for the timing of H1 — the major post-LGM episode of IRD deposition in the North Atlantic — but most terminate this event before or at the time of MWP-1A^{58,62} and the relationship between the two events remains unclear. MWP-1A is followed by the Younger Dryas stadial during which the thermohaline circulation did slow down and large-scale cooling occurred, but not to the same extent as occurred during the MIS-3 events. It is possible that the astronomical forcing of climate was now sufficiently influential to disrupt the earlier high-frequency cyclic behaviour and, as for the early part of the glacial cycle, was able to exert its dominating influence on climate.

More questions than answers

During the Pleistocene epoch of Earth history, glaciations have dominated and interglacials occurred for less than 10% of the time. The

current interglacial interval has persisted for about as long as the duration of earlier interglacials and a return to stadial conditions should not be unanticipated if the past is any guide to the future; sea-level and climate fluctuations during the last interglacial may well provide an interesting analogue for an eventual termination of the present interglacial.

The causes of the climate fluctuations that repeatedly built up and destroyed ice sheets remain unclear. Sea-level fluctuations during the last glacial cycle have responded to the dominant oscillations in insolation, with periodicities of ~ 40 kyr and ~ 20 kyr (Fig. 2). But the relative amplitudes and phase relationships show no consistent match. The onset of the maximum in sea level preceded the peak in Northern Hemisphere summer insolation in the last interglacial^{31,44,47,48}, whereas for the present interglacial the insolation peak preceded the sea-level maximum. The insolation minima at ~ 140 and ~ 20 kyr ago correspond to the two associated glacial maxima, but at ~ 115 kyr ago, for example, insolation was even lower, without a correspondingly lower value in sea level.

The timing of the penultimate glacial maximum has not yet been adequately constrained by quantitative chronological data, but the rise in sea level seems to start at about the time of the solar minimum. Likewise, the end of the LGM coincides with an insolation minimum. The stadials MIS-5d and -5b, and possibly MIS-4, also occur at the time of minima in insolation. This is perhaps the most consistent correlation between the two functions throughout this cycle.

The lack of an overall correlation may be related to the influence on climate of other parameters, in addition to the summer insolation at 65°N . For example, a higher-latitude insolation index may be more appropriate or, more significantly, Southern Hemisphere insolation may also be important, and Antarctic ice may respond independent of Northern Hemisphere changes or with a time lag⁶². Various attempts at defining an insolation index that leads to a more clear-cut correlation have so far proved unsuccessful^{99,100}.

The structure of the sea-level curve shows greater high-frequency fluctuations than does the insolation index. This sea-level evidence, along with other proxy indicators of climate, points to a multitude of interrelated forces and feedbacks that are more important at some periods than others (see also the commentary in this issue by Kump, pages 188–190). It seems that during the early part of the glacial cycle, insolation controls the glacial oscillations and sea level but, starting at about MIS-4, feedback mechanisms become increasingly important. By the time ice sheets have become long-term features of the Northern Hemisphere landscape, reaching the continental shelves and becoming unstable, the subtle interplay between Milankovitch cycles, ice-sheet dynamics and shifts in ocean circulation seems to drive the climate system. At ~ 20 kyr ago, and possibly during the penultimate glacial maximum, the oscillations in insolation once again take over and break the higher-frequency feedback cycle. Insolation now dominates, notwithstanding brief oscillations back into ice-age conditions caused by the feedback processes, as evidenced by a rapid sea-level oscillation during the transition from MIS-6 to MIS-5³⁷ and the short-lived, stationary sea level during the Younger Dryas cold period⁶⁴. □

doi:10.1038/nature01089

- Mix, A. C. & Ruddiman, W. F. Oxygen isotope analyses and Pleistocene ice volumes. *Quat. Res.* **21**, 1–20 (1984).
- Shackleton, N. J. Oxygen isotope analyses and Pleistocene temperatures re-assessed. *Nature* **215**, 15–17 (1967).
- Raymo, M. E. In *Start of a Glacial* (eds Kukla, G. J. & Went, E.) 207–223 (Springer, Berlin, 1992).
- Haug, G. H. & Tiedeman, R. Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermohaline circulation. *Nature* **393**, 673–676 (1998).
- Shackleton, N. J. Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the North Atlantic region. *Nature* **307**, 620–623 (1984).
- Shackleton, N. J. & Opdyke, N. D. Oxygen isotope and palaeomagnetic stratigraphy of equatorial Pacific core V28-238: oxygen isotope temperatures and ice volumes on a 10^3 year scale and 10^4 year scale. *Quat. Res.* **3**, 39–55 (1973).
- Broecker, W. S. & van Donk, J. Insolation changes, ice volumes and the ^{18}O record in deep-sea cores. *Rev. Geophys. Space Phys.* **8**, 169–198 (1970).
- Chappell, J. Geology of coral terraces, Huon Peninsula, New Guinea: a study of Quaternary tectonic

- movements and sea-level changes. *Bull. Geol. Soc. Am.* **85**, 553–570 (1974).
9. Pirazolli, P. A. *et al.* Quaternary raised coral-reef terraces on Sumba Island, Indonesia. *Science* **252**, 1834–1836 (1991).
 10. Stirling, C. H. *et al.* Orbital forcing of the marine isotope stage 9 interglacial. *Science* **291**, 290–293 (2001).
 11. Broecker, W. S. & Denton, G. H. The role of ocean-atmosphere reorganizations in glacial cycles. *Geochim. Cosmochim. Acta* **53**, 2465–2501 (1989).
 12. McManus, J. F., Oppo, D. W. & Cullen, J. L. A 0.5-million-year record of millennial-scale climate variability in the North Atlantic. *Science* **283**, 971–975 (1999).
 13. MacAyeal, D. R. Binge/purge oscillations of the Laurentide ice sheets as a cause of the North Atlantic's Heinrich events. *Paleoceanography* **8**, 775–784 (1993).
 14. Alley, R. B. & Clark, P. U. The deglaciation of the northern hemisphere: a global perspective. *Annu. Rev. Earth Planet. Sci.* **27**, 149–182 (1999).
 15. Hay, W. W. The cause of the Late Cenozoic Northern Hemisphere glaciations: a climate change enigma. *Terra Nova* **4**, 305–311 (1992).
 16. Berger, A. Milankovitch theory and climate. *Rev. Geophys.* **26**, 624–657 (1988).
 17. Emiliani, C. Pleistocene temperatures. *J. Geol.* **63**, 538–578 (1955).
 18. Hays, J. D., Imbrie, J. & Shackleton, N. J. Variations in the earth's orbit: pacemaker of the ice ages. *Science* **194**, 1121–1132 (1976).
 19. Duplessy, J. C., Chenouard, L. & Villa, F. Weyl's theory of glaciation supported by isotopic study of Norwegian core K11. *Science* **188**, 1208–1209 (1975).
 20. Imbrie, J. *et al.* On the structure and origin of major glaciation cycles, 2. The 100,000-year cycle. *Paleoceanography* **8**, 699–735 (1993).
 21. Berger, A., Loutre, M. F. & Lascar, J. Stability of the astronomical frequencies over the earth's history for paleoclimatic studies. *Science* **255**, 560–566 (1992).
 22. Imbrie, J. *et al.* On the structure and origin of major glaciation cycles, 1. Linear responses to Milankovitch forcing. *Paleoceanography* **7**, 701–738 (1992).
 23. Imbrie, J., Berger, A. & Shackleton, N. J. In *Global Changes in the Perspective of the Past* (eds Eddy, J. A. & Oeschger, H.) 263–277 (Wiley, Chichester, 1993).
 24. Imbrie, J. & Imbrie, J. Z. Modelling the climatic responses to orbital variations. *Science* **207**, 943–953 (1980).
 25. Clark, P. U. & Pollard, D. Origin of the middle Pleistocene transition by ice sheet erosion of regolith. *Paleoceanography* **13**, 1–9 (1998).
 26. Muller, R. A. & MacDonald, G. J. F. Glacial cycles and astronomical forcing. *Science* **277**, 215–218 (1997).
 27. Shackleton, N. J. The 100,000-year ice-age cycle identified and found to lag temperature, carbon dioxide, and orbital eccentricity. *Science* **289**, 1897–1902 (2000).
 28. Gallup, C. D., Edwards, R. L. & Johnson, R. G. The timing of high sea levels over the past 200,000 years. *Science* **263**, 796–800 (1994).
 29. Bard, E., Antonioli, F. & Silenzi, S. Duration and timing of the penultimate interglacial sea level highstand: implications for the astronomical theory of paleoclimate. *Earth Planet. Sci. Lett.* **196**, 135–146 (2002).
 30. Galewsky, J., Silver, E. A., Gallup, C. D., Edwards, R. L. & Potts, D. C., Foredeep tectonics and carbonate platform dynamics in the Huon Gulf, Papua New Guinea. *Geology* **24**, 819–822 (1996).
 31. Stirling, C. H., Esat, T. M., Lambeck, K. & McCulloch, M. T. Timing and duration of the Last Interglacial: evidence for a restricted interval of widespread coral reef growth. *Earth Planet. Sci. Lett.* **160**, 745–762 (1998).
 32. Rohling, E. J. *et al.* Magnitudes of the sea-level lowstands of the past 500,000 years. *Nature* **394**, 162–165 (1998).
 33. Ferland, M. A., Roy, P. S. & Murray-Wallace, C. V. Glacial lowstand deposits on the outer continental shelf of southeastern Australia. *Quat. Res.* **44**, 294–299 (1995).
 34. Chappell, J. & Shackleton, N. J. Oxygen isotopes and sea level. *Nature* **324**, 137–140 (1986).
 35. Meselella, K., Matthews, R. K., Broecker, W. S. & Thurber, D. L. The astronomical theory of climate change: Barbados data. *J. Geol.* **77**, 250–274 (1969).
 36. Chappell, J. *et al.* Reconciliation of late Quaternary sea levels derived from coral terraces at Huon Peninsula with deep sea oxygen isotope records. *Earth Planet. Sci. Lett.* **141**, 227–236 (1996).
 37. Esat, T. M., McCulloch, M. T., Chappell, J., Pillans, B. & Omura, A. Rapid fluctuations in sea level recorded at Huon Peninsula during the Penultimate Deglaciation. *Science* **283**, 197–201 (1999).
 38. Yokoyama, Y., Esat, T. M., Lambeck, K. & Fifield, L. K. Last ice age millennial scale climate changes recorded in Huon Peninsula corals. *Radiocarbon* **42**, 383–401 (2000).
 39. Stein, M. *et al.* TIMS U-series dating and stable isotopes of the last interglacial event in Papua New Guinea. *Geochim. Cosmochim. Acta* **57**, 2541–2554 (1993).
 40. Lambeck, K. & Chappell, J. Sea level change through the last glacial cycle. *Science* **292**, 679–686 (2001).
 41. Yokoyama, Y., Esat, T. M. & Lambeck, K. Coupled climate and sea-level changes deduced from Huon Peninsula coral terraces of the last ice age. *Earth Planet. Sci. Lett.* **193**, 579–587 (2001).
 42. Chappell, J. C. Sea-level changes forced ice breakouts in the Last Glacial cycle: new results from coral terraces. *Quat. Sci. Rev.* **21**, 1229–1240 (2002).
 43. Imbrie, J. *et al.* In *Milankovitch and Climate* (eds Berger, A. *et al.*) 269–305 (Reidel, Dordrecht, 1984).
 44. Lambeck, K. & Nakada, M. Constraints on the age and duration of the last interglacial period and on sea-level variations. *Nature* **357**, 125–128 (1992).
 45. Potter, E.-K. *et al.* Timing and magnitude of oxygen isotope stage 5a and 5c sea level oscillations at Barbados: implications for the melting history of the Laurentide ice sheet during OIS 5 and 6. *Eos Trans. AGU* **82**(47), Fall Meet. Suppl., Abstract PP12A-0467 (2001).
 46. Gallup, C. D., Cheng, H., Taylor, F. W. & Edwards, R. L. Direct determination of the timing of sea level change during termination II. *Science* **295**, 310–313 (2002).
 47. Winograd, I. J. *et al.* Continuous 500,000 year climate record from vein calcite in Devils Hole, Nevada. *Science* **258**, 255–260 (1992).
 48. Henderson, G. M. & Slowey, N. C., Evidence from U-Th dating against northern hemisphere forcing of the penultimate deglaciation. *Nature* **404**, 61–65 (2000).
 49. Ludwig, K. R. *et al.* Sea level records at ~80ka from tectonically stable platforms: Florida and Bermuda. *Geology* **24**, 211–214 (1996).
 50. Toscano, M. A. & Lundberg, J. Submerged Late Pleistocene reefs on the tectonically-stable S.E. Florida margin: high-precision geochronology, stratigraphy, resolution of Substage 5a sea-level elevation, and orbital forcing. *Quat. Sci. Rev.* **18**, 753–767 (1999).
 51. Denton, G. H. & Hughes, T. J. (eds) *The Last Great Ice Sheets* (Wiley, New York, 1980).
 52. Lundqvist, J. In *Glacial Stratigraphy, Engineering Geology and Earth Construction* (ed. Kauranne, K.) (Geol. Surv. Finland Special Paper 15) 43–59 (Geol. Surv. Finland, 1992).
 53. Helmens, K. F., Räsänen, M. E., Johansson, P. W., Jungner, H. & Korjonen, K. The last interglacial-glacial cycle in NE Scandinavia: a nearly continuous record from Sokli (Finnish Lapland). *Quat. Sci. Rev.* **19**, 1605–1623 (2000).
 54. Mangerud, J. Ice sheet limits in Norway and on the Norwegian continental shelf. *Quat. Sci. Rev.* (submitted).
 55. Svendsen, J. I. *et al.* Maximum extent of the Eurasian ice sheets in the Barents and Kara Sea region during the Weichselian. *Boreas* **28**, 234–242 (1999).
 56. Clark, P. U. *et al.* Initiation and development of the Laurentide and Cordilleran ice sheets following the Last Interglaciation. *Quat. Sci. Rev.* **12**, 79–114 (1993).
 57. Lund, D. C. & Mix, A. C. Millennial-scale deep water oscillations: reflections of the North Atlantic in the deep Pacific from 10 to 60 ka. *Paleoceanography* **13**, 10–19 (1998).
 58. Bond, G. *et al.* Evidence for massive discharges of icebergs into the North Atlantic ocean during the last glacial period. *Nature* **360**, 245–249 (1992).
 59. Bond, G. *et al.* Correlations between climate records from North Atlantic sediments and Greenland ice. *Nature* **365**, 143–147 (1993).
 60. Kanfoush, S. L. *et al.* Millennial-scale instability of the Antarctic ice sheet during the last glaciation. *Science* **288**, 1815–1818 (2000).
 61. Heinrich, H. Origin and consequences of cyclic ice rafting in the northeast Atlantic Ocean during the last 130,000 years. *Quat. Res.* **29**, 143–152 (1988).
 62. Chapman, M. R., Shackleton, N. J. & Duplessy, J.-C., Sea surface temperature variability during the last glacial-interglacial cycle assessing the magnitude and pattern of climate change in the North Atlantic. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **157**, 1–25 (2000).
 63. Broecker, W. S. Massive iceberg discharges as triggers for global climate change. *Nature* **372**, 421–424 (1994).
 64. Fawcett, P. J., Agustodottir, A. M., Alley, R. B. & Shuman, C. A. The Younger Dryas termination and the North Atlantic deep water formation: insights from climate model simulations and Greenland ice cores. *Paleoceanography* **12**, 23–38 (1997).
 65. Alley, R. B. Icing the North Atlantic. *Nature* **392**, 335–337 (1998).
 66. Kitagawa, H. & van der Plicht, J., Atmospheric radiocarbon calibration to 45,000 yr B.P.: late glacial fluctuations and cosmogenic isotope production. *Science* **279**, 1187–1190 (1998).
 67. Hughen, K. A., Southon, J. R., Lehman, S. J. & Overpeck, J. T. Synchronous radiocarbon and climate shifts during the last deglaciation. *Science* **290**, 1951–1954 (2000).
 68. Broecker, W. S. & Hemming, S. Climate swings come into focus. *Science* **294**, 2308–2309 (2001).
 69. Dowdeswell, J. A., Elverhøi, A., Andrews, J. T. & Hebbeln, D. Asynchronous deposition of ice-rafted layers in the Nordic seas and North Atlantic Ocean. *Nature* **400**, 348–351 (1999).
 70. Shemesh, A., Burckle, L. H. & Hays, J. D. Meltwater input to the Southern Ocean during the Last Glacial Maximum. *Science* **266**, 1542–1544 (1994).
 71. Bond, G. C. & Lotti, R. Iceberg discharges into the North Atlantic on millennial time scales during the Last Glaciation. *Science* **267**, 1005–1010 (1995).
 72. Mix, A. C., Bard, E. & Schneider, R. Environmental processes of the ice age land, oceans, glaciers (EPILOG). *Quat. Sci. Rev.* **20**, 627–657 (2001).
 73. Yokoyama, Y., Lambeck, K., De Deckker, P., Johnston, P. & Fifield, L. K. Timing of the Last Glacial Maximum from observed sea-level minima. *Nature* **406**, 713–716 (2000). [Published correction in *Nature*, **412**, 99 (2001).]
 74. Milne, G. A., Mitrovica, J. X. & Schrag, D. P. Estimating past continental ice volume from sea-level data. *Quat. Sci. Rev.* **21**, 361–376 (2002).
 75. Lambeck, K., Yokoyama, Y. & Purcell, A., Into and out of the Last Glacial Maximum: sea-level change during oxygen isotope stages 3 and 2. *Quat. Sci. Rev.* **21**, 343–360 (2002).
 76. Olsen, L., Svein, H. & Bergström, B. Rapid adjustments of the western part of the Scandinavian ice sheet during the Mid and Late Weichselian—a new model. *Norw. J. Geol.* **81**, 93–118 (2001).
 77. Sejrup, H. P. *et al.* Quaternary glaciations in southern Fennoscandia: evidence from southwestern Norway and the northern North Sea. *Quat. Sci. Rev.* **19**, 667–685 (2000).
 78. Ukkonen, P., Lunkka, J. P., Jungner, H. & Donner, J. New radiocarbon dates from Finnish mammoths indicating large ice-free areas in Fennoscandia during the Middle Weichselian. *J. Quat. Sci.* **14**, 711–714 (1999).
 79. Tushingham, A. M. & Peltier, W. R. ICE-3G: a new global model of Late Pleistocene deglaciation based upon geophysical predictions of postglacial sea-level change. *J. Geophys. Res.* **96**, 4497–4523 (1991).
 80. Lambeck, K., Smither, C. & Johnston, P. Sea-level change, glacial rebound and mantle viscosity for northern Europe. *Geophys. J. Int.* **134**, 647–651 (1998).
 81. Vorren, T. O., Vorren, K.-D., Alm, T., Gullimssen, S. & Løvlie, R. The last deglaciation (20,000–11,000 B.P.) on Andøya, northern Norway. *Boreas* **17**, 41–77 (1988).
 82. Licciardi, J. M., Clark, P. U., Jenson, J. W. & MacAyeal, D. R. Deglaciation of a soft-bed Laurentide ice sheet. *Quat. Sci. Rev.* **17**, 427–448 (1998).
 83. Miller, G. H. *et al.* The Goldilocks dilemma: big ice, little ice, or “just-right” ice in the Eastern Canadian Arctic. *Quat. Sci. Rev.* **21**, 33–48 (2002).
 84. Lambeck, K. Limits on the areal extent of the Barents Sea ice sheet in Late Weichselian time. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **12**, 41–51 (1996).
 85. Bennike, O., Björck, S. & Lambeck, K. Estimates of South Greenland late-glacial ice limits from a new relative sea level curve. *Earth Planet. Sci. Lett.* **197**, 171–186 (2002).
 86. Zwartz, D., Bird, M., Stone, J. & Lambeck, K. Holocene sea-level change and ice-sheet history in the Vestfold Hills, East Antarctica. *Earth Planet. Sci. Lett.* **155**, 131–145 (1998).
 87. CLIMAP Project Members. The surface of the ice age Earth. *Science* **191**, 1131–1137 (1976).
 88. Lambeck, K., Yokoyama, Y., Johnston, P. & Purcell, A. Global ice volumes at the Last Glacial Maximum and early Late glacial. *Earth Planet. Sci. Lett.* **181**, 513–527 (2000).
 89. Elliot, M. *et al.* Millennial-scale iceberg discharges in the Irmering Basin during the last glacial period: relationship with the Heinrich events and environmental settings. *Paleoceanography* **13**, 433–446 (1998).
 90. Shemesh, A., Burckle, L. H. & Hays, J. D. Meltwater input to the Southern Ocean during the Last Glacial Maximum. *Science* **266**, 1542–1544 (1994).
 91. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).

92. Stoffers, P. & Ross, D. A. Late Pleistocene and Holocene sedimentation in the Persian Gulf-Gulf of Oman. *Sedim. Geol.* **23**, 181–208 (1979).
93. Fairbanks, R. G. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. *Nature* **342**, 637–642 (1989).
94. Bard, E., Hamelin, B. & Fairbanks, R. G. U-Th ages obtained by mass spectrometry in corals from Barbados: sea level during the past 130,000 years. *Nature* **346**, 456–458 (1990).
95. Bard, E. *et al.* Deglacial sea-level record from Tahiti corals and the timing of global meltwater discharge. *Nature* **382**, 241–244 (1996).
96. Hanebuth, T., Statteger, K. & Grootes, P. M. Rapid flooding of the Sunda Shelf: a late-glacial sea-level record. *Science* **288**, 1033–1035 (2000).
97. Clark, P. U. *et al.* Origin of the first global meltwater pulse following the last glacial maximum. *Paleoceanography* **11**, 563–577 (1996).
98. Clark, P. U., Mitrovica, J. X., Milne, G. A. & Tamisiea, M. E. Sea-level fingerprinting as a direct test for the source of global meltwater pulse 1A. *Science* **295**, 2438–2441 (2002).
99. Johnson, R. G. Major northern hemisphere deglaciation caused by a moisture deficit 140 ka. *Geology* **19**, 686–689 (1991).
100. Shaffer, J. A., Cervený, R. S. & Dom, R. I. Radiation windows as indicators of an astronomical influence on the Devil's Hole chronology. *Geology* **24**, 1017–1020 (1996).

Acknowledgements

This research was supported by the Australian National University and by the Swedish Research Council. We thank S. Björck for constructive comments and M. E. Raymo and G. H. Haug for providing the data sets for Fig. 1.

Ocean circulation and climate during the past 120,000 years

Stefan Rahmstorf

Potsdam Institute for Climate Impact Research, PO Box 601203, 14412 Potsdam, Germany

Oceans cover more than two-thirds of our blue planet. The waters move in a global circulation system, driven by subtle density differences and transporting huge amounts of heat. Ocean circulation is thus an active and highly nonlinear player in the global climate game. Increasingly clear evidence implicates ocean circulation in abrupt and dramatic climate shifts, such as sudden temperature changes in Greenland on the order of 5–10 °C and massive surges of icebergs into the North Atlantic Ocean — events that have occurred repeatedly during the last glacial cycle.

The world ocean is one of the main constituents of the climate system and affects climate in a multitude of ways. Its sheer size is obvious: 71% of the Earth is covered by ocean, so that most of the solar radiation received at the Earth's surface goes into the ocean and warms the surface waters. As a result of its heat capacity and circulation, the ocean has the ability to both store and redistribute this heat before it is released to the atmosphere (much of it in form of latent heat, that is, water vapour) or radiated back into space.

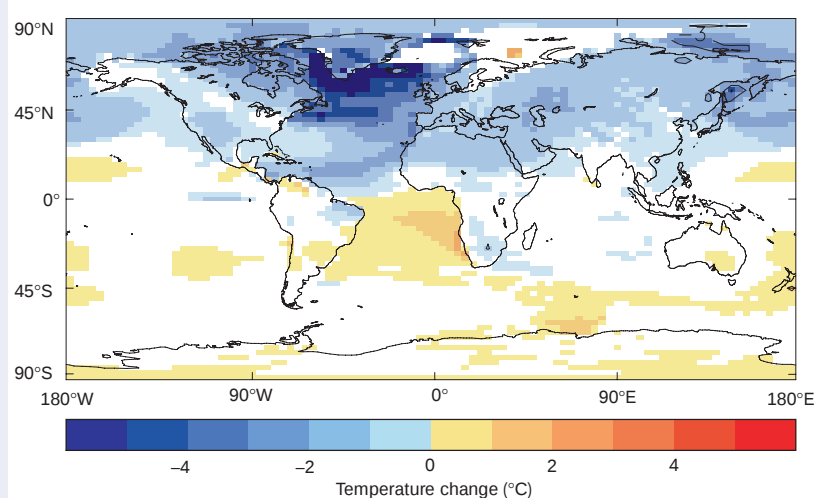
The heat storage effect is most apparent on the seasonal timescale. The mid-latitude temperature range between summer and winter is typically around 8 °C over the ocean and at the coast, whereas this range is up to several tens of degrees in the continental interiors (see Figure 2.1 in ref. 1, which contours the observed seasonal temperature range).

A corresponding figure for the temperature deviation from the zonal mean (Fig. 1 in ref. 2) gives an indication of

the effect of ocean heat transport on surface temperatures, with warm anomalies over the three main regions of deep-water formation of the world ocean: the northern North Atlantic, the Ross Sea and the Weddell Sea. These are key areas for the thermohaline circulation of the world ocean (see Box 1), where surface waters after releasing heat to the atmosphere reach a critical density and sink. Clearly not all deviations from zonal mean conditions are due to ocean circulation, but the magnitude of the warm anomaly over the northern North Atlantic (~10 °C) is in agreement with estimates³ and simulations of climate models^{4–7} of the effect of ocean heat transport (Fig. 1).

In addition to its heat storage and transport effects, the ocean can influence the Earth's heat budget by its sea-ice cover, which changes the planetary albedo and can thus affect the steady-state global-mean temperature. Sea ice also acts as an effective thermal blanket, insulating the ocean from the overlying atmosphere. This is so effective that in a typical ice-covered sea more than half of the air–sea heat

Figure 1 Changes in surface air temperature caused by a shutdown of North Atlantic Deep Water (NADW) formation in a current ocean–atmosphere circulation model. Note the hemispheric seesaw (Northern Hemisphere cools while the Southern Hemisphere warms) and the maximum cooling over the northern Atlantic. In this particular model (HadCM3)⁷, the surface cooling resulting from switching off NADW formation is up to 6 °C. It is further to the west compared with most models, which tend to put the maximum cooling near Scandinavia. This probably depends on the exact location of deep-water formation (an aspect not well represented in current



coarse-resolution models) and on the sea-ice distribution in the models, as ice-margin shifts act to amplify the cooling. The largest air temperature cooling is thus greater than the largest sea surface temperature (SST) cooling. The latter is typically around 5 °C and roughly corresponds to the observed SST difference between the northern Atlantic and Pacific at a given latitude. In most models, maximum air temperature cooling ranges from 6 °C to 11 °C in annual mean; the effect is generally stronger in winter.

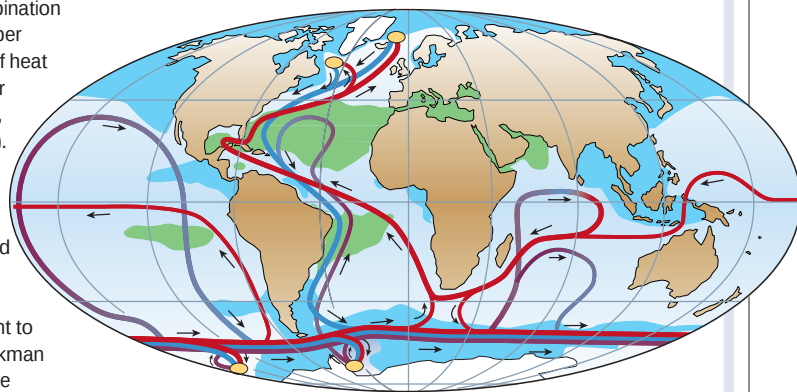
Box 1

Some key facts about ocean circulation

The large-scale ocean circulation can be thought of as a combination of currents driven directly by winds (mostly confined to the upper several hundred metres of the sea), currents driven by fluxes of heat and freshwater across the sea surface and subsequent interior mixing of heat and salt (the so-called thermohaline circulation), and tides (driven by the gravitational pull of the Moon and Sun). These driving mechanisms interact in nonlinear ways (since all currents change the heat and salt distribution) so that no unique decomposition exists. Nevertheless the distinction is useful, particularly when changes in wind or in surface heat and freshwater fluxes are considered for their effects on the circulation.

An important way in which wind-driven currents are thought to lead to climatic changes is through their effect on upwelling (Ekman divergence) near coasts and the Equator, changing sea surface temperatures. This mechanism plays a part in the El Niño/Southern Oscillation cycle. The thermohaline circulation is most interesting for its highly nonlinear response to changes in surface freshwater forcing⁸⁸, allowing large changes in heat transport to occur (see Box 2). Tides are relevant to the climate system because they form one of the main sources of turbulent energy (in addition to that provided by the wind) to mix the ocean⁸⁹.

A highly simplified cartoon of the global thermohaline circulation (sometimes called 'conveyor belt') is shown in the figure above (modified from the original by Broecker). Near-surface waters (red lines) flow towards three main deep-water formation regions (yellow ovals) — in the northern North Atlantic, the Ross Sea and the Weddell Sea — and recirculate at depth (deep currents shown in blue, bottom currents in purple; green shading indicates salinity above 36‰, blue shading indicates salinity below 34‰). A recent estimate of the rate of deep-water formation is 15 ± 2 Sv ($1 \text{ Sv} = 10^6 \text{ m}^3 \text{ s}^{-1}$) in the North Atlantic and 21 ± 6 Sv in the Southern Ocean⁹⁰. Northward heat transport into the northern Atlantic peaks at 1.3 ± 0.1 PW ($1 \text{ PW} = 10^{15} \text{ W}$) in the subtropics⁹⁰; this heat transport warms the northern Atlantic regional air temperatures by up to 10°C



over the ocean with the effect declining inland.

Little is currently known about present-day natural variability of this circulation (see ref. 91 for a review), or about the effects of such variability on surface climate. Variations of the Atlantic thermohaline circulation on timescales of several decades are found in many coupled climate models, with a typical amplitude of a few sverdrup; they are probably damped oscillations driven by stochastic variations in surface fluxes (that is, weather variability)⁹². Good observational time series of integral measures of this circulation are lacking, although some data suggest that such decadal variability also exists in nature, and is correlated with the North Atlantic Oscillation (NAO)^{93,94}. The NAO also seems to orchestrate the location and intensity of deep convection in the northern Atlantic⁹⁵. Lack of data makes it hard to establish whether a longer-term trend in the circulation exists, although there is intriguing evidence for trends in temperature and salinity^{96,97} that may indicate a gradual weakening of the overflow from the Nordic Seas into the Atlantic in recent decades.

exchange occurs through patches of open water (leads) that make up around 10% of the surface area.

Finally, the ocean affects the climate system not only by being part of the planetary energy cycle, but also by participating in the biogeochemical cycles and exchanging gases with the atmosphere, thus influencing its greenhouse gas content. For example, the ocean contains about fifty times more carbon than the atmosphere, and theories seeking to explain the lower concentrations of atmospheric carbon dioxide that prevailed during glacial times invariably invoke changes in the oceanic carbon sink, either through physical or biological mechanisms (the so-called 'biological pump').

Rather than providing a general overview of the ocean's role in the climate system, which is a subject matter for textbooks, I focus here on the role of ocean circulation changes in major climate changes during the past 120,000 years, since the Eemian interglacial. This is a period for which palaeoclimatic data of relatively good global coverage and dating are available. The emphasis is on presenting physical ideas and concepts for understanding these climate changes; more-detailed reviews of the palaeoclimatic data can be found elsewhere (see, for example, refs 8–10). This is a highly dynamical research field with rapid progress, but not yet a generally accepted and established theory. Controversies remain over many issues, and the interpretation I have attempted here is subjective and will probably turn out to be partly wrong.

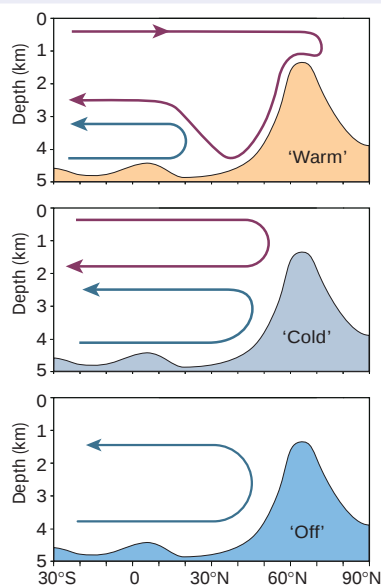
Reconstructing past ocean circulation

Analysis of sediment cores and corals provides a wealth of information on past ocean circulation, and clearly shows that it has

undergone significant changes during the past 120,000 years. Reconstructions of past ocean temperatures can be derived, for example, from species abundances of fossil plankton, from organic geochemistry (using alkenone unsaturation indices), from trace-metal ratios (Sr/Ca, U/Ca or Mg/Ca) in corals or calcite shells, and to some extent from oxygen isotopes. Using multiple proxies, information on salinity can also be reconstructed. This constrains the distribution and properties of water masses; information on flow rates is harder to obtain. Indirect evidence for ventilation rates comes from the distribution of isotopes such as ^{13}C (ref. 11) or radiochemical tracers¹², and from the radiocarbon content of the atmosphere. In some locations, the grain size of sediments yields information on the speed of local bottom currents¹³, or density gradients have been reconstructed to give information on the geostrophic current component¹⁴. Although there is still much discussion on the interpretation and error margins of each data type, and in some cases proxy data seem to give contradictory results, an increasingly consistent picture is emerging.

Time-slice compilations suggest that at different times, three distinct circulation modes have prevailed in the Atlantic^{15,16} (Fig. 2). These have been labelled the stadial mode, interstadial mode and Heinrich mode (based on their occurrence during stadial and interstadial phases of glacial climate and during Heinrich events), or the cold, warm and off mode (based on their physical characteristics in the North Atlantic). In the interstadial mode, North Atlantic Deep Water (NADW) formed in the Nordic Seas, in the stadial mode it formed in the subpolar open North Atlantic (that is, south of Iceland), whereas

Figure 2 Schematic of the three modes of ocean circulation that prevailed during different times of the last glacial period. Shown is a section along the Atlantic; the rise in bottom topography symbolizes the shallow sill between Greenland and Scotland. North Atlantic overturning is shown by the red line, Antarctic bottom water by the blue line.



in the Heinrich mode NADW formation all but ceased and waters of Antarctic origin filled the deep Atlantic basin. This grouping of the data in three distinct modes is a somewhat subjective interpretation. However, it is clear that latitude shifts of convection (between the Nordic Seas and the region south of Iceland) have occurred^{16,17}, and that at certain times (for example, during Heinrich events) NADW formation was interrupted^{11,18}. There is also firm evidence now for a link between these changes in ocean circulation and changes in surface climate (argued in more detail in refs 8, 19).

Modelling past climate and ocean changes

Numerical models of the climate system are essential in the formation and exploration of quantitative hypotheses about the dynamics of climate changes; the system is too complex to be understood by heuristic arguments or analytical calculations. Numerical models incorporate and combine our knowledge about many individual physical processes in a quantitative way. Obviously, knowledge about these processes is incomplete and often inaccurate, and each model is a compromise as to how many processes are included, at what level of complexity and with what resolution²⁰, given limited computer and human resources. A critical appraisal of what can be learnt from a particular (necessarily imperfect) model experiment thus involves not only looking at the result, but also understanding exactly how it was obtained. For this reason, non-specialists sometimes suspect that models are either notoriously wrong or 'can be tuned to do anything'. (In fact, 'tuning' to determine the optimal values for certain model

parameters is an essential part of constructing a good model; a set of rules for good tuning practice is proposed in ref. 21.)

Nevertheless, models have now reached a level where useful and fairly realistic simulations of many aspects of palaeoclimate have become possible, so that a quantitative understanding of key mechanisms and feedbacks in past climate changes is emerging. On the other hand, palaeoclimatic reconstructions of past climatic forcings and the resulting changes in atmospheric and oceanic conditions are now advanced enough to provide a challenging test bed for the performance of climate models. This is an important credibility test for models that are also used for estimating the effects of anthropogenic climate forcing from increasing concentrations of greenhouse gases.

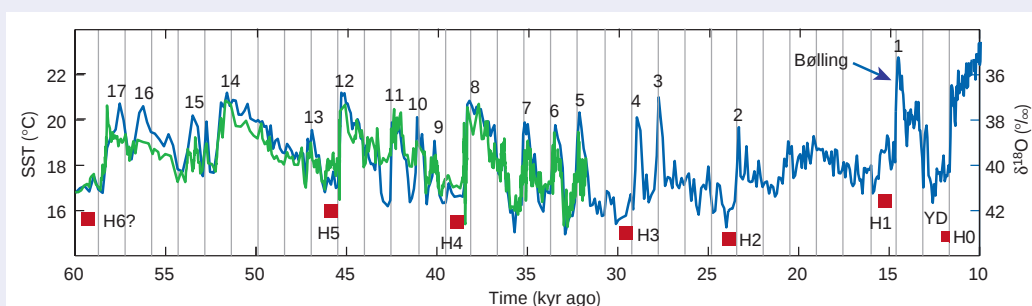
A landmark was reached with the first simulations of a radically different climate, that of the Last Glacial Maximum (LGM), with coupled ocean–atmosphere models from prescribed orbital, CO₂ and continental ice-sheet forcing^{22–25}. The huge computing requirements of this task, resulting from the long timescale of adjustment of the ocean circulation (several thousand years), were overcome in different ways: by using fast models of intermediate complexity^{22,23}, by studying only the initial adjustment to the forcing²⁴, or by brute force, running the model for over a year on a supercomputer²⁵. These models confirm the result of the much cheaper, atmosphere-only simulations of glacial climate (see, for example, ref. 26) that, given these forcings, the high albedo of the continental ice sheets and the low CO₂ concentrations are the dominant factors leading to a global cooling. In addition, the coupled models predict the state of the ocean circulation and the effect of oceanic changes on surface climate. For example, two of the models^{22,25} obtained a southward shift of NADW formation in glacial climate, as is suggested by sediment data^{16,17}.

Glacial inception

The first of the major climatic changes considered here is the transition from the Eemian interglacial to the beginning of glacial climate, which occurred between 120,000 years (120 kyr) and 115 kyr ago. Data for the Eemian climate are too scarce to build a reliable picture, but global simulations of Eemian climate together with local palaeodata suggest it may have been around 1 °C warmer (global annual mean) compared with the modern pre-industrial climate²⁷, with particularly warm temperatures in Northern Hemisphere summers. Sea-level reconstructions²⁸ show that climate moved rapidly from this state into the last glacial, reaching almost half of the glacial-maximum ice volume within a few thousand years (see review in this issue by Lambeck *et al.*, pages 199–206). The challenge of understanding this shift has become known as the 'glacial inception problem'.

The cause for this climate shift must be the Milankovich cycles of the Earth's orbit around the Sun, which are believed to be the ultimate forcing for the glacial cycles of the past 2 million years. At 115 kyr ago, summer insolation at high northern latitudes was up to 40 W m⁻² less than at present. Simple, nonlinear conceptual models²⁹ are able to

Figure 3 Temperature reconstructions from ocean sediments and Greenland ice. Proxy data from the subtropical Atlantic⁸⁶ (green) and from the Greenland ice core GISP2 (ref. 87; blue) show several Dansgaard–Oeschger (D/O) warm events (numbered). The timing of Heinrich events is marked in red.



Grey lines at intervals of 1,470 years illustrate the tendency of D/O events to occur with this spacing, or multiples thereof.

reproduce the observed glaciations from this forcing (according to these, the next glaciation can be expected in ~30 kyr from now).

A number of climate models have been used to study glacial inception even without incorporating a continental ice-sheet model, based on the concept that snow cover that persists throughout the summer would eventually grow into an ice sheet. Discussion has focused on the conditions under which sufficient perennial snow cover can be achieved. This has turned out to be difficult in atmosphere-only models with present-day sea surface temperatures and vegetation cover. A reduction in high-latitude forest cover greatly increases the albedo after snowfall, leading to a positive snow-albedo feedback that helps glacial inception³⁰. Lower high-latitude sea surface temperatures (caused by the insolation change) are also clearly conducive to perennial snow cover³¹. Recently, a realistic simulation of glacial inception in terms of actual ice-sheet growth has been achieved in an Earth system model of intermediate complexity that includes a continental ice-sheet model (R. Calov & A. Ganopolski, in preparation).

Does a weakening in Atlantic Ocean circulation have a role in glacial inception? There are no palaeoclimatic data showing that NADW formation slowed at this time. Model simulations that include a dynamical ocean model achieve glacial inception with only minor changes in ocean circulation (ref. 31; and R. Calov & A. Ganopolski, in preparation) — changes that are too small to be important in glacial inception in these models. The reverse theory³², namely that a warm North Atlantic could have induced ice-sheet growth by enhanced moisture supply, goes against our knowledge of glacier mass balance: glaciers grow when climate is cold, not warm and moist.

Dansgaard–Oeschger events

Dansgaard–Oeschger (D/O) events (Figs 3, 4) are perhaps the most pronounced climate changes that have occurred during the past 120 kyr. They are not only large in amplitude, but also abrupt (irrespective of whether one follows a physical definition of abruptness³³ or takes it to mean ‘in less than 30 years’⁸). In the Greenland ice cores, D/O events start with a rapid warming by 5–10 °C within at most a few decades, followed by a plateau phase with slow cooling lasting several centuries, then a more rapid drop back to cold stadial conditions. The events are not local to Greenland (Fig. 3); a comprehensive review of spatial coverage (for events during marine isotope stage 3, 59–29 kyr ago) is given by Voelker *et al.*¹⁰ who list 183 sites, most of which clearly show these events (Fig. 4). Amplitudes are largest in the North Atlantic region, and many Southern Hemisphere sites, especially those in the South Atlantic, reveal a hemispheric ‘see-saw’ effect (cooling while the north is warming). Alley *et al.*³⁴ have shown that these events have curious statistical properties: the waiting time between two consecutive events is often around 1,500 years, with further preferences around 3,000 and 4,500 years (Fig. 3), which suggests a stochastic resonance³⁵ process at work.

Several ideas have been advanced to explain D/O events, most of which involve the thermohaline circulation of the Atlantic. The first of these was probably the idea of thermohaline circulation bistability³⁶; NADW formation is active during the warm phases (interstadials), whereas it is shut off during cold phases (stadials), and some outside trigger causes mode switches between these two stable states. This idea is based on the bistability of the circulation for modern climate in models (Broecker³⁶ cites the bistability found in Stommel’s³⁷ classic model; see Box 2). However, this theory is at odds with more recent sediment data showing NADW formation active during stadials^{12,15} and shut down only during or after Heinrich events^{11,18}.

A second idea is that of internal oscillations in the volume transport of the thermohaline circulation. Broecker’s salt oscillator³⁸ is based on the (challenged³⁹) notion that the Atlantic thermohaline circulation balances the net atmospheric freshwater export from the Atlantic basin. A weakening of the circulation would thus lead to a salinity build-up in the Atlantic, strengthening the circulation again. A variation of this idea involves the surrounding continental ice

sheets: a strong circulation warms the northern Atlantic and melts surrounding ice, which leads to meltwater runoff and weakens the circulation again^{36,40}. These were conceptual ideas, but circulation models are also able to show several types of internal oscillations in thermohaline flow (without ice sheets) under certain forcing conditions⁴¹. The relevance of such model oscillations for the real ocean is open to debate, and a problem of all internal-oscillation theories for D/O events is to explain the waiting-time statistics found by Alley and co-workers.

A third idea is that of latitude shifts of convection³ between Nordic Seas and the mid-latitude open Atlantic Ocean. Based originally on sediment data, this idea has found strong support in model simulations showing that such a mechanism can explain many observed features of D/O events⁴², including the three-phase time evolution,

Box 2

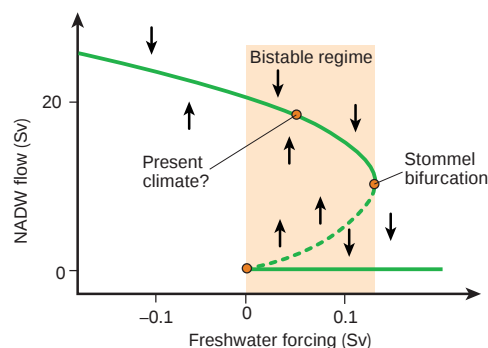
Stability and nonlinearity of the thermohaline circulation

The thermohaline circulation is thermally driven: highest surface densities in the world ocean are reached where water is coldest, in spite of the lower salt content there compared with the warmer tropical and subtropical regions. Nevertheless, the influence of salinity is important and is the main cause of the nonlinearity of the system. Salinity is involved in a positive feedback. Higher salinity in the deep-water formation area enhances the circulation, and the circulation in turn transports higher salinity waters into the deep-water formation regions (which tend to be regions of net precipitation, that is, freshwater would accumulate and surface salinity would drop if the circulation stopped). This leads to two possible equilibrium states, with and without North Atlantic Deep Water (NADW) formation⁴. This was first described in a classic paper by Stommel³⁷ with the help of a simple box model.

The stability properties are illustrated in the diagram below, which plots the strength of the thermohaline circulation as a function of the freshwater input into the North Atlantic. The simple presentation shows the bistable regime and a saddle-node bifurcation point where the circulation breaks down (for a more detailed discussion, see ref. 88).

This salt-transport feedback is not the only feedback rendering the system nonlinear. The convective mixing process at high latitudes is itself a highly nonlinear, self-sustaining process, which at least in models can lead to multiple stable convection patterns^{3,98}. Together, these two positive feedback mechanisms allow two types of transitions between distinct circulation modes: on/off switches of NADW formation, and shifts in the location of convection. These two mechanisms are crucial in attempts to explain glacial climate changes.

The thermohaline circulation takes several thousand years to reach full equilibrium. The transient response to a change in forcing can therefore deviate substantially from the equilibrium solutions and is in many cases more linear⁹⁹.



spatial pattern and hemispheric see-saw. In this mechanism, the rapid warming phase results from a northward intrusion of warm Atlantic waters into the Nordic Seas, the plateau phase is the 'warm mode' of Atlantic Ocean circulation (see above), which gradually weakens over several centuries, and the final cooling phase marks the end of deep-water formation in the Nordic Seas. Some trigger is required to start the event, the exact nature of which remains unknown. However, because this is a threshold mechanism it lends itself naturally to stochastic resonance, with random climate variability plus a weak external cycle in freshwater forcing (for example, driven by a solar cycle^{43,44}) combining to cross the critical threshold^{45,46}.

Finally, the tropical driver hypothesis^{47,48} does not involve changes in thermohaline circulation, but suggests that D/O-style temperature shifts in Greenland may be caused by shifts in the atmospheric planetary-wave pattern, controlled remotely from the tropics. This is based on the strong control that tropical sea surface temperatures exert over global atmospheric heat-transport patterns in present climate, but a more specific and quantitative explanation for D/O events building on this idea is yet to emerge.

It is possible that several of these basic physical ideas work together in D/O events. For example, shifts in convection latitude could be caused by changes in atmospheric freshwater transport controlled partly from the tropics.

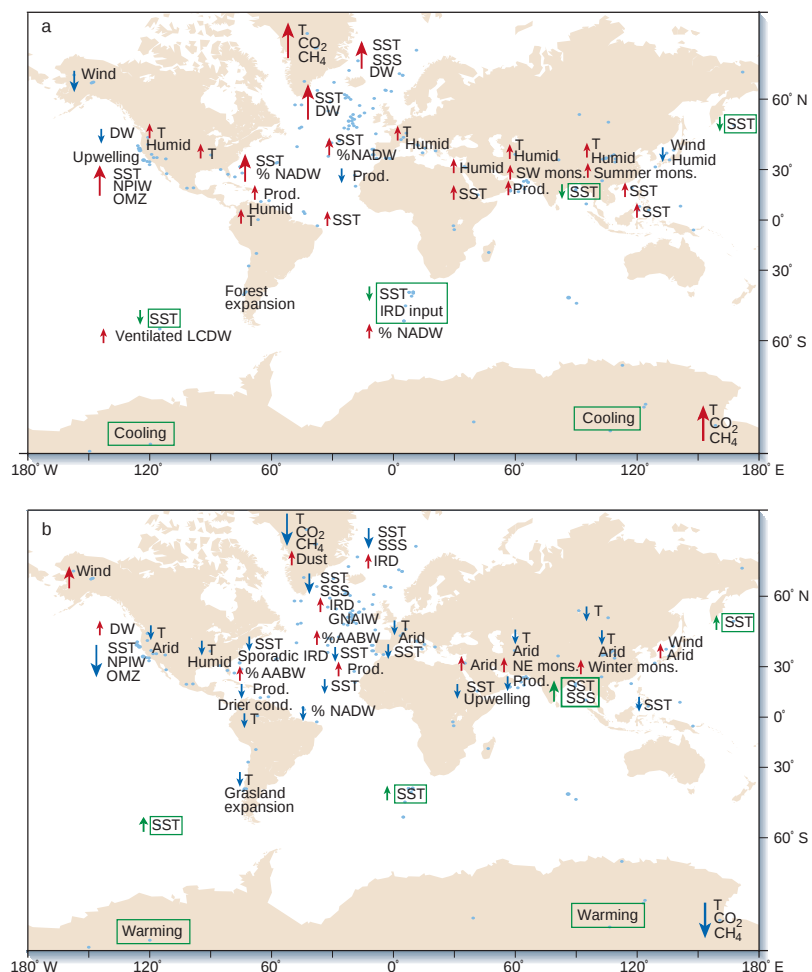
Heinrich events

Heinrich events are the second major type of climatic event that occurred mostly in the latter half of the last glacial. They are

characterized by distinct layers in North Atlantic sediments^{49,50}, spaced at irregular intervals of the order of 10,000 years. Sediments in these layers are so coarse that they can only have been transported out into the ocean by icebergs; hence, they are referred to as ice-rafted debris. The thickness of these layers, decreasing from several metres in the Labrador Sea down to a few centimetres in the eastern Atlantic, strongly suggests that Heinrich events are massive episodic iceberg discharges from the Laurentide ice sheet through Hudson Strait, with up to 10% of the ice sheet sliding into the ocean^{51–53}. A highly plausible explanation is that the ice sheet grew to a critical height where it became unstable, and a major surge could then start spontaneously or be triggered by a small perturbation^{54,55}. Sediment data show that NADW formation ceased or was at least strongly reduced during Heinrich events^{11,15,56}, and models consistently show that this is to be expected^{6,42,57–60} given the reduction in surface water density associated with such a large freshwater release (up to 0.1 Sv; ref. 53).

The climatic consequences of Heinrich events thus probably consist of the superposition of two effects: the direct effect of the ice-sheet surge, leading to a lowered ice sheet and higher sea level, and the effect of the subsequent breakdown of the Atlantic thermohaline circulation. The climate signature of Heinrich events differs from D/O events in several ways. In Greenland, stadials were equally cold with or without Heinrich events. Further south around the Atlantic, however, Heinrich events manifest as clear cold intervals with even larger amplitude than D/O warmings^{61–63}. This pattern can be explained if the 'latitude shift of convection' theory of D/O events is

Figure 4 Overview of palaeoclimatic proxy data¹⁰ characterizing warm phases (top) and cold phases (bottom) during marine oxygen isotope stage 3 (MIS-3; 59–29 kyr ago, compare with Fig. 3). Red arrows (blue arrows) indicate trends, that is, warmer (colder), more (less) or increased (lowered). Green text and arrows indicate trends opposite to the general climate conditions. Abbreviations: T, temperature; SST, sea surface temperature; SSS, sea surface salinity; mons., monsoon; prod., productivity; cond., conditions; IRD, ice-rafted debris; OMZ, oxygen minimum zone. Water masses are labelled as follows: DW, Deep Water; NADW, North Atlantic Deep Water; AABW, Antarctic Bottom Water; NPIW, North Pacific Intermediate Water; LCDW, Lower Circumpolar Deep Water; GNAIW, Glacial North Atlantic Intermediate Water.



correct: during stadials, the ocean is in the 'cold mode' with the warm Atlantic current stopping too far south to warm Greenland, so that shutting it down has no effect there⁴² (but it does further south).

The data further show that at most Antarctic sites, Heinrich events are associated with warming that is stronger than that during other stadials⁶⁴. This is the bipolar see-saw (or 'sea-saw') effect^{65,66}, resulting from the reduced interhemispheric heat transport by the ocean; the effect of a shutdown of NADW formation is greater than that of a latitude shift⁴². There are in fact two see-saw effects in operation: in addition to the temperature see-saw, many models indicate there is a deep-water formation see-saw, which will lead to enhanced Southern Ocean deep-water formation if NADW formation is reduced, and vice versa. As well as enhancing the see-saw response in temperature, this mechanism gives a possible role to changes in the Southern Ocean deep-water source in affecting northern Atlantic climate⁶⁷.

D/O and Heinrich events are not unrelated. First, each Heinrich event is followed by a particularly warm D/O event; successive D/O events tend to get progressively cooler until the next Heinrich event (this sequence of D/O events is sometimes referred to as a Bond cycle). This could simply be a consequence of the Laurentide ice sheet growing gradually in height between Heinrich events.

Second, Heinrich events apparently always occur during cold stadials and not in the warm phase of D/O events⁵¹. This suggests that ice-sheet instability does not occur at random, but is helped by some climatic trigger, possibly a temperature or sea-level change⁶⁸; there is also evidence for smaller precursor events⁶⁹. The issue of a possible trigger mechanism, which may also synchronize discharges from separate ice sheets⁶⁸, is one of the important, currently open research questions surrounding Heinrich events.

Deglaciation and the Younger Dryas event

The end of the last ice age and the transition to the Holocene is the last first-order global climatic change on record. Since then, climate has been relatively warm and stable, providing a conducive environment for the development of human civilization. A rise in (ice-volume-equivalent) sea level by 130 m between 19 kyr and 7 kyr ago marks the rapid vanishing of the glacial ice sheets²⁸. Many puzzles surround the complex sequence of events that occurred during deglaciation, and three key factors need to be considered: the changes in insolation (due to the Milankovich cycles) which must have initiated deglaciation, the rise in atmospheric CO₂ levels providing a strong global warming feedback, and changes in ocean circulation.

Surface warming started around 17–20 kyr ago in Antarctica and proceeded approximately synchronously with the rise in atmospheric CO₂ and global sea level. Northern records such as the Greenland ice cores, however, show a very different deglaciation history. A consistent and plausible (although tentative) explanation can be advanced if we assume (as for the D/O and Heinrich events discussed above) that these northern sites are dominated by the state of the Atlantic thermohaline circulation, which went through some significant changes during deglaciation, in part because of the influx of meltwater from the shrinking ice sheets.

According to this explanation, warming from glacial-maximum conditions is initiated by changes in northern insolation⁷⁰ (summer insolation increases in high latitudes of the Northern Hemisphere by about 30 W m⁻² between 24 and 12 kyr ago). The carbon cycle responds almost synchronously by releasing CO₂ to the atmosphere⁷¹, which reinforces and globalizes the warming (together with other greenhouse gases, primarily water vapour), and the ice sheets start to melt. Greenland, however, remains cold (although some warming begins at a similar time as in Antarctica when the two regions are viewed on a common timescale⁷²), as meltwater influx and Heinrich event 1 tend to keep the Atlantic Ocean in cold circulation mode⁷³. Greenland then warms abruptly at 14.6 kyr ago in the Bølling warming, owing to a northward shift in ocean circulation (that is, D/O event 1). If we believe the chronology derived from the

Greenland ice core GISP2, this follows almost exactly 9 kyr after D/O 2, thus fitting a multiple of the 1,500-year cycle.

Synchronous with the Bølling warming is a small cooling in Antarctica — the Antarctic cold reversal — which interrupts the general trend towards warming there, and which represents the characteristic see-saw response to the change in Atlantic Ocean circulation associated with D/O 1. A major inflow of meltwater⁷⁴ into the ocean (meltwater pulse 1A) is registered shortly after the strong northern warming. This could be a consequence of the Bølling warming, assuming most of this meltwater originated from the northern ice sheets (a question that is debated). But the meltwater did not immediately close down NADW formation, perhaps because it did indeed originate mostly in the Southern Hemisphere, or perhaps because the North Atlantic was at this time in the vigorous warm interstadial mode, which is relatively insensitive to freshwater forcing. D/O 1 finally ended (as did the previous D/O events), giving way to the Younger Dryas (YD) cold event (warming accelerates in Antarctica at this point). Finally, at 11.5 kyr ago, the YD ends with an abrupt warming that might be called D/O 0, almost exactly 3 kyr after D/O 1, again fitting the 1,500-year pattern. This warming is also followed by another major meltwater pulse (pulse 1B), but this time the North Atlantic remains in the warm circulation mode, which is stable in warmer climates and prevails in the Holocene.

Model simulations with the CLIMBER-2 model show that this is a feasible scenario that can be reproduced from prescribed insolation, CO₂, ice-sheet and meltwater forcing. However, the sequence and timing of events is particularly sensitive to the details of the freshwater forcing, which are poorly known. Efforts have been made to estimate the history of meltwater flux from different outlets and link it to ocean circulation and climate changes⁷³, but additional data are needed to obtain a more robust representation, and the freshwater forcing may never be known accurately enough to allow truly deterministic modelling.

The Younger Dryas event seems to be special in a number of ways. Because of the high meltwater influx at this time, NADW formation probably stopped^{58,75,76}, as during Heinrich events. Nevertheless, it seems hard to reconcile the fact that the Younger Dryas event is almost as cold as previous Heinrich events during glacial-maximum conditions with the already elevated CO₂ level in the atmosphere (over 240 p.p.m.) and reduced inland ice volume. Furthermore, there is increasing evidence from New Zealand⁷⁷ and South America^{78,79} that the Younger Dryas event was accompanied by a global re-advance of ice, which is also reflected in a temporary halt of sea-level rise²⁸. The Younger Dryas event may thus be more than a change in ocean circulation; a global forcing causing cooling could be involved, possibly of solar origin⁸⁰.

A final northern cooling in the history of deglaciation is a short event occurring 8,200 years ago, which has also been linked to a meltwater-induced weakening of the thermohaline circulation⁸¹.

El Niño/Southern Oscillation

In present-day climate, the strongest mode of natural climate variability is the El Niño/Southern Oscillation (ENSO). It is a coupled ocean–atmosphere mode centred on the tropical Pacific, with a variable period of 3–7 years and worldwide ecological and societal impacts due to its effect on the global atmospheric circulation. Annually banded corals provide a unique opportunity to determine whether this mode has also been in operation during different climatic states of the past, as they record climatic information at up to monthly resolution in the chemistry of their skeletons as they grow⁸². As a result of tectonic uplift, fossil coral reefs from past climatic periods can be found on exposed terraces at some sites, such as the Huon Peninsula of Papua New Guinea (see review in this issue by Lambeck *et al.*, pages 199–206).

Coral data from different time segments show convincingly that ENSO variability prevailed in very different climates, including glacial times and the Eemian interglacial⁸². The amplitude seems to

have varied, however, with particularly weak ENSO variations during the mid-Holocene (6.5 kyr ago) and the early glacial (112 kyr ago) and the strongest ENSO during modern times. First attempts to simulate the effect of Milankovich cycles on ENSO variations using a simple model suggest that the precession cycle could directly alter ENSO intensity by zonally asymmetric heating of the equatorial Pacific⁸³. But comparison with data shows that this cannot be the only effect⁸², and both more data and further simulations with more comprehensive models are needed to understand ENSO variations through time.

Outlook

The study of climate variations over the past 120,000 years has reached a state where palaeoclimatic data provide increasingly reliable information on the driving forces and the responses of the climate system, and where distinct climatic events such as glaciation, deglaciation, D/O events or Heinrich events can be characterized in terms of their spatial patterns and evolution over time. Understanding the mechanisms behind these climatic changes has moved beyond speculation to specific, testable hypotheses backed up by quantitative simulations.

It has become clear that the climate system is sensitive to forcing and responds with large and often abrupt changes in surface conditions. The role of the ocean circulation is that of a highly nonlinear amplifier of climatic changes. Many issues are still controversial and unresolved, both in terms of the data (for example, whether the late-glacial glacier advance in New Zealand and South America is synchronous with the Younger Dryas cold event in the north) and in terms of the mechanisms (for example, whether Younger Dryas cooling is caused by a meltwater-induced shutdown of NADW formation). But progress has been rapid, and the potential exists to resolve many of these issues in the coming decade or so by collecting more data, refining the analysis methods and improving models.

A better understanding of the carbon cycle remains one of the main challenges; the ocean has a crucial role in this cycle, one that could not be discussed here owing to space limitations. Reconstructions^{71,84} and modelling⁸⁵ of carbon cycle changes can provide useful constraints on ocean circulation changes, and understanding the glacial–interglacial changes in atmospheric CO₂ concentration remains an elusive central piece in the climate puzzle. □

doi:10.1038/nature01090

- Gill, A. E. *Atmosphere–Ocean Dynamics* (Academic, San Diego, 1982).
- Rahmstorf, S. & Ganopolski, A. Long-term global warming scenarios computed with an efficient coupled climate model. *Clim. Change* **43**, 353–367 (1999).
- Rahmstorf, S. Rapid climate transitions in a coupled ocean–atmosphere model. *Nature* **372**, 82–85 (1994).
- Manabe, S. & Stouffer, R. J. Two stable equilibria of a coupled ocean–atmosphere model. *J. Clim.* **1**, 841–866 (1988).
- Ganopolski, A. *et al.* CLIMBER-2: a climate system model of intermediate complexity. Part II: Model sensitivity. *Clim. Dynam.* **17**, 735–751 (2001).
- Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature* **378**, 145–149 (1995).
- Vellinga, M. & Wood, R. A. Global climatic impacts of a collapse of the Atlantic thermohaline circulation. *Clim. Change* **54**, 251–267 (2002).
- Clark, P. U., Pisias, N. G., Stocker, T. F. & Weaver, A. J. The role of the thermohaline circulation in abrupt climate change. *Nature* **415**, 863–869 (2002).
- Mix, A. C., Bard, E. & Schneider, R. Environmental processes of the ice age: land, oceans, glaciers (EPILOG). *Quat. Sci. Rev.* **20**, 627–657 (2001).
- Voelker, A. H. L. *et al.* Global distribution of centennial-scale records for marine isotope stage (MIS) 3: a database. *Quat. Sci. Rev.* **21**, 1185–1214 (2002).
- Elliot, M., Labeyrie, L. & Duplessy, J.-C. Changes in North Atlantic deep-water formation associated with the Dansgaard–Oeschger temperature oscillations (60–10 ka). *Quat. Sci. Rev.* **21**, 1153–1165 (2002).
- Yu, E.-F., Francois, R. & Bacon, M. P. Similar rates of modern and last-glacial ocean thermohaline circulation inferred from radiochemical data. *Nature* **379**, 689–694 (1996).
- Bianchi, G. G. & McCave, I. N. Holocene periodicity in North Atlantic climate and deep-ocean flow south of Iceland. *Nature* **397**, 515–517 (1999).
- Lynch-Stiegitz, J., Curry, W. B. & Slowey, N. Weaker Gulf Stream in the Florida Straits during the Last Glacial Maximum. *Nature* **402**, 644–648 (1999).
- Sarnthein, M. *et al.* Changes in east Atlantic deepwater circulation over the last 30,000 years: eight time slice reconstructions. *Paleoceanography* **9**, 209–267 (1994).
- Alley, R. B. & Clark, P. U. The deglaciation of the Northern Hemisphere: a global perspective. *Annu. Rev. Earth Planet. Sci.* **27**, 149–182 (1999).

- Oppo, D. & Lehman, S. J. Mid-depth circulation of the subpolar North Atlantic during the Last Glacial Maximum. *Science* **259**, 1148–1152 (1993).
- Keigwin, L. D. & Lehman, S. J. Deep circulation change linked to Heinrich event 1 and Younger Dryas in a mid-depth North Atlantic core. *Paleoceanography* **9**, 185–194 (1994).
- Bond, G. *et al.* Correlations between climate records from North Atlantic sediments and Greenland ice. *Nature* **365**, 143–147 (1993).
- Claussen, M. *et al.* Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Clim. Dynam.* **18**, 579–586 (2002).
- Petoukhov, V. *et al.* CLIMBER-2: a climate system model of intermediate complexity. Part I: Model description and performance for present climate. *Clim. Dynam.* **16**, 1–17 (2000).
- Ganopolski, A., Rahmstorf, S., Petoukhov, V. & Claussen, M. Simulation of modern and glacial climates with a coupled global model of intermediate complexity. *Nature* **391**, 351–356 (1998).
- Weaver, A. J., Eby, M., Fanning, A. F. & Wiebe, E. C. Simulated influence of carbon dioxide, orbital forcing and ice sheets on the climate of the Last Glacial Maximum. *Nature* **394**, 847–853 (1998).
- Bush, A. B. G. & Philander, S. G. H. The role of ocean–atmosphere interactions in tropical cooling during the Last Glacial Maximum. *Science* **279**, 1341–1344 (1998).
- Hewitt, C. D., Broccoli, A. J., Mitchell, J. F. B. & Stouffer, R. J. A coupled model study of the last glacial maximum: was part of the North Atlantic relatively warm? *Geophys. Res. Lett.* **28**, 1571–1574 (2001).
- Webb, R. S., Rind, D. H., Lehman, S. J., Healy, R. J. & Sigman, D. Influence of ocean heat transport on the climate of the Last Glacial Maximum. *Nature* **385**, 695–699 (1997).
- Kubatzki, C., Montoya, M., Rahmstorf, S., Ganopolski, A. & Claussen, M. Comparison of the last interglacial climate simulated by a coupled global model of intermediate complexity and an AOGCM. *Clim. Dynam.* **16**, 799–814 (2000).
- Lambeck, K. & Chappell, J. Sea level change through the last glacial cycle. *Science* **292**, 679–686 (2001).
- Paillard, D. Glacial cycles: toward a new paradigm. *Rev. Geophys.* **39**, 325–346 (2001).
- Gallimore, R. G. & Kutzbach, J. E. Role of orbitally induced changes in tundra area in the onset of glaciation. *Nature* **381**, 503–505 (1996).
- Khodri, M. *et al.* Simulating the amplification of orbital forcing by ocean feedbacks in the last glaciation. *Nature* **410**, 570–574 (2001).
- Gildor, H. & Tziperman, E. A sea ice climate switch mechanism for the 100-kyr glacial cycles. *J. Geophys. Res.* **106**, 9117–9133 (2001).
- Rahmstorf, S. in *Encyclopedia of Ocean Sciences* (eds Steele, J., Thorpe, S. & Turekian, K.) 1–6 (Academic, London, 2001).
- Alley, R. B., Anandakrishnan, S. & Jung, P. Stochastic resonance in the North Atlantic. *Paleoceanography* **16**, 190–198 (2001).
- Gammaitoni, L., Hanggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–287 (1998).
- Broecker, W. S., Peteet, D. M. & Rind, D. Does the ocean–atmosphere system have more than one stable mode of operation? *Nature* **315**, 21–26 (1985).
- Stommel, H. Thermohaline convection with two stable regimes of flow. *Tellus* **13**, 224–230 (1961).
- Broecker, W. S., Bond, G., Klas, M., Bonani, G. & Wolfi, W. A salt oscillator in the glacial North Atlantic? 1. The concept. *Paleoceanography* **5**, 469–477 (1990).
- Rahmstorf, S. On the freshwater forcing and transport of the Atlantic thermohaline circulation. *Clim. Dynam.* **12**, 799–811 (1996).
- Birchfield, G. E., Wang, H. & Rich, J. J. Century/millennium internal climate variability: an ocean–atmosphere–continental icesheet model. *J. Geophys. Res.* **99**, 12459–12470 (1994).
- Winton, M. & Sarachik, E. S. Thermohaline oscillations induced by strong steady salinity forcing of ocean general circulation models. *J. Phys. Oceanogr.* **23**, 1389–1410 (1993).
- Ganopolski, A. & Rahmstorf, S. Rapid changes of glacial climate simulated in a coupled climate model. *Nature* **409**, 153–158 (2001).
- Bond, G. *et al.* Persistent solar influence on North Atlantic climate during the holocene. *Science* **294**, 2130–2136 (2001).
- Van Geel, B. *et al.* The role of solar forcing upon climate change. *Quat. Sci. Rev.* **18**, 331–338 (1999).
- Rahmstorf, S. & Alley, R. B. Stochastic resonance in glacial climate. *Eos* **83**, 129–135 (2002).
- Ganopolski, A. & Rahmstorf, S. Abrupt glacial climate changes due to stochastic resonance. *Phys. Rev. Lett.* **88**, 038501–1–038501–4 (2002).
- Clement, A. C. & Cane, M. A. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 363–371 (Am. Geophys. Union, Washington DC, 1999).
- Cane, M. A. & Clement, A. C. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 373–383 (Am. Geophys. Union, Washington DC, 1999).
- Heinrich, H. Origin and consequences of cyclic ice rafting in the northeast Atlantic Ocean during the past 130,000 years. *Quat. Res.* **29**, 143–152 (1988).
- Hemming, S. R., Bond, G. C., Broecker, W. S., Sharp, W. D. & Klas-Mendelson, M. Evidence from Ar-40/Ar-39 ages of individual hornblende grains for varying Laurentide sources of iceberg discharges 22,000 to 10,500 yr BP. *Quat. Res.* **54**, 372–383 (2000).
- Bond, G. *et al.* Evidence for massive discharges of icebergs into the North Atlantic ocean during the last glacial. *Nature* **360**, 245–249 (1992).
- Andrews, J. T. Abrupt changes (Heinrich events) in late Quaternary North Atlantic marine environments: a history and review of data and concepts. *J. Quat. Sci.* **13**, 3–16 (1998).
- Chappell, J. Sea level changes forced ice breakouts in the last glacial cycle: new results from coral terraces. *Quat. Sci. Rev.* **21**, 1229–1240 (2002).
- MacAyeal, D. R. Binge/purge oscillations of the Laurentide ice sheet as a cause of the North Atlantic's Heinrich events. *Paleoceanography* **8**, 775–784 (1993).
- Clark, P. U., Alley, R. B. & Pollard, D. Northern hemisphere ice sheet influences on global climate change. *Science* **286**, 1104–1111 (1999).
- Keigwin, L. D., Curry, W. B., Lehman, S. J. & Johnsen, S. The role of the deep ocean in North Atlantic climate change between 70 and 130 kyr ago. *Nature* **371**, 323–326 (1994).
- Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean. *Nature* **378**, 165–167 (1995).
- Maier-Reimer, E., Mikolajewicz, U., Wooster, W. & Yáñez-Arancibia, A. in *Oceanography* (ed. Ayala-Castañares, A.) 87–100 (National Autonomous University (UNAM) Press, Mexico, 1989).
- Stocker, T. F. & Wright, D. G. Rapid transitions of the ocean's deep circulation induced by changes in surface water fluxes. *Nature* **351**, 729–732 (1991).
- Weaver, A. J. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 285–300 (Am. Geophys. Union, Washington DC, 1999).

61. Paillard, D. & Cortijo, E. A simulation of the Atlantic meridional circulation during Heinrich event 4 using reconstructed sea surface temperatures and salinities. *Paleoceanography* **14**, 716–724 (1999).
62. Cacho, I. et al. Dansgaard-Oeschger and Heinrich event imprints in the Alboran Sea paleotemperatures. *Paleoceanography* **14**, 698–705 (1999).
63. Bard, E., Rostek, F., Turon, J.-L. & Gendreau, S. Hydrological impact of Heinrich events in the subtropical Northeast Atlantic. *Science* **289**, 1321–1324 (2000).
64. Blunier, T. et al. Asynchrony of Antarctic and Greenland climate change during the last glacial period. *Nature* **394**, 739–743 (1998).
65. Crowley, T. J. North Atlantic deep water cools the Southern Hemisphere. *Paleoceanography* **7**, 489–497 (1992).
66. Stocker, T. F. The seesaw effect. *Science* **282**, 61–62 (1998).
67. Seidov, D., Haupt, B. J., Barron, E. J. & Maslin, M. in *The Oceans and Rapid Climate Change: Past, Present, and Future* (eds Seidov, D., Haupt, B. J. & Maslin, M.) 147–167 (Am. Geophys. Union, Washington DC, 2001).
68. Bond, G. C. & Lotti, R. Iceberg discharges into the North Atlantic on millennial time scales during the last glaciation. *Science* **267**, 1005–1010 (1995).
69. Bond, G. in *Mechanisms of Global Climate Change at Millennial Time Scales* (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 35–58 (Am. Geophys. Union, Washington DC, 1999).
70. Alley, R. B., Brook, E. J. & Anandakrishnan, S. A northern lead in the orbital band: north-south phasing of ice-age events. *Quat. Sci. Rev.* **21**, 431–441 (2002).
71. Monnin, E. et al. Atmospheric CO₂ concentrations over the last glacial termination. *Science* **291**, 112–114 (2001).
72. Blunier, T. & Brook, E. J. Timing of millennial-scale climate change in Antarctica and Greenland during the last glacial period. *Science* **291**, 109–112 (2001).
73. Clark, P. U. et al. Freshwater forcing of abrupt climate change during the last glaciation. *Science* **293**, 283–287 (2001).
74. Fairbanks, R. G. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. *Nature* **342**, 637–642 (1989).
75. Fanning, A. F. & Weaver, A. J. Temporal-geographical meltwater influences on the North Atlantic conveyor: implications for the Younger Dryas. *Paleoceanography* **12**, 307–320 (1997).
76. Manabe, S. & Stouffer, R. Coupled ocean-atmosphere model response to freshwater input: comparison to Younger Dryas event. *Paleoceanography* **12**, 321–336 (1997).
77. Denton, G. H. & Hendy, C. H. Younger Dryas advance of Franz Josef Glacier in the Southern Alps of New Zealand. *Science* **264**, 1434–1437 (1994).
78. Hajdas, I., Bonani, G., Moreno, P. I. & Ariztegui, D. Precise radiocarbon dating of a Younger Dryas-age cooling in mid-latitude South America. A step towards inter-hemispheric climate linkage. *Quat. Res.* (in the press).
79. Moreno, P. I., Jacobson, G. L., Lowell, T. V. & Denton, G. H. Interhemispheric climate links revealed by a late-glacial cooling episode in southern Chile. *Nature* **409**, 804–808 (2001).
80. Renssen, H., Van Geel, B., Van der Plicht, J. & Magny, M. Reduced solar activity as a trigger for the start of the Younger Dryas? *Quat. Int.* **68–71**, 373–383 (2001).
81. Renssen, H., Goosse, H., Fichefet, T. & Campin, J.-M. The 8.2 kyr BP event simulated by a global atmosphere-sea-ice-ocean model. *Geophys. Res. Lett.* **28**, 1567–1570 (2001).
82. Tudhope, A. W. et al. Variability in the El Niño–Southern Oscillation through a glacial-interglacial cycle. *Science* **291**, 1511–1517 (2001).
83. Clement, A., Seager, R. & Cane, M. A. Suppression of El Niño during the mid-holocene by changes in the earth's orbit. *Paleoceanography* **15**, 731–737 (2000).
84. Muscheler, R., Beer, J., Wagner, G. & Finkel, R. G. Changes in deep-water formation during the Younger Dryas event inferred from ¹⁰Be and ¹⁴C records. *Nature* **408**, 567–570 (2000).
85. Marchal, O. et al. Modelling the concentration of atmospheric CO₂ during the Younger Dryas climate event. *Clim. Dynam.* **15**, 341–354 (1999).
86. Sachs, J. P. & Lehman, S. J. Subtropical North Atlantic temperatures 60,000 to 30,000 years ago. *Science* **286**, 756–759 (1999).
87. Groote, P. M., Stuiver, M., White, J. W. C., Johnsen, S. & Jouzel, J. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature* **366**, 552–554 (1993).
88. Rahmstorf, S. The thermohaline ocean circulation—a system with dangerous thresholds? *Clim. Change* **46**, 247–256 (2000).
89. Munk, W. & Wunsch, C. Abyssal recipes II: energetics of wind and tidal mixing. *Deep Sea Res.* **45**, 1977–2010 (1998).
90. Ganachaud, A. & Wunsch, C. Improved estimates of global ocean circulation, heat transport and mixing from hydrographic data. *Nature* **408**, 453–457 (2000).
91. Rahmstorf, S. in *Beyond El Niño: Decadal and Interdecadal Climate Variability* (ed. Navarra, A.) 309–332 (Springer, Berlin, 1999).
92. Delworth, T., Manabe, S. & Stouffer, R. J. Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model. *J. Clim.* **6**, 1993–2011 (1993).
93. Bacon, S. Decadal variability in the outflow from the Nordic Seas to the deep Atlantic Ocean. *Nature* **394**, 871–874 (1998).
94. Marshall, J. & et al. North Atlantic climate variability: phenomena, impacts and mechanisms. *Int. J. Clim.* **21**, 1863–1898 (2001).
95. Dickson, R. R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term co-ordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
96. Dickson, R. R. et al. Rapid freshening of the deep North Atlantic over the past four decades. *Nature* **410**, 832–837 (2001).
97. Hansen, B., Turrell, W. R. & Østerhus, S. Decreasing overflow from the Nordic seas into the Atlantic Ocean through the Faroe Bank channel since 1950. *Nature* **411**, 927–930 (2001).
98. Lenderink, G. & Haarsma, R. J. Variability and multiple equilibria of the thermohaline circulation, associated with deep water formation. *J. Phys. Oceanogr.* **24**, 1480–1493 (1994).
99. Stouffer, R. J. & Manabe, S. Response of a coupled ocean-atmosphere model to increasing atmospheric carbon dioxide: sensitivity to the rate of increase. *J. Clim.* **12**, 2224–2237 (1999).

Acknowledgements

This manuscript has benefited greatly from the advice of A. Ganopolski, R. Alley, G. Bond and M. Cane, and from the lively discussions within the National Oceanic and Atmospheric Administration's Panel on Abrupt Climate Change.

A satellite view of aerosols in the climate system

Yoram J. Kaufman*, Didier Tanré† & Olivier Boucher†

*NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA (e-mail: kaufman@climate.gsfc.nasa.gov)

†Laboratoire d'Optique Atmosphérique, Université de Lille/CNRS, Villeneuve d'Ascq, France

Anthropogenic aerosols are intricately linked to the climate system and to the hydrologic cycle. The net effect of aerosols is to cool the climate system by reflecting sunlight. Depending on their composition, aerosols can also absorb sunlight in the atmosphere, further cooling the surface but warming the atmosphere in the process. These effects of aerosols on the temperature profile, along with the role of aerosols as cloud condensation nuclei, impact the hydrologic cycle, through changes in cloud cover, cloud properties and precipitation. Unravelling these feedbacks is particularly difficult because aerosols take a multitude of shapes and forms, ranging from desert dust to urban pollution, and because aerosol concentrations vary strongly over time and space. To accurately study aerosol distribution and composition therefore requires continuous observations from satellites, networks of ground-based instruments and dedicated field experiments. Increases in aerosol concentration and changes in their composition, driven by industrialization and an expanding population, may adversely affect the Earth's climate and water supply.

During the last century, the Earth's surface temperature increased by 0.6 °C, reaching the highest levels in the last millennium¹. This rapid temperature change is attributed to a shift of less than 1% (ref. 2) in the energy balance between absorption of incoming solar radiation and emission of thermal radiation from the Earth system. Among the different agents of climate change, anthropogenic greenhouse gases and aerosols have the larger roles¹. Whereas greenhouse gases reduce the emission of thermal radiation to space, thereby warming the surface, aerosols mainly reflect and absorb solar radiation (the aerosol direct effect) and modify cloud properties (the aerosol indirect effect), cooling the surface. These impacts on the radiation balance are very different and therefore require different research approaches.

Greenhouse gases, such as carbon dioxide and methane, have a lifetime of up to 100 years in the atmosphere and a rather homogeneous distribution around the globe; this is in contrast to the heterogeneous spatial and temporal distribution of tropospheric aerosols, which results from their short lifetime of about a week^{1,3}. As a consequence, the global increase in the CO₂ concentration of 1–2 p.p.m. per year was measured half a century ago using a single ground-based instrument⁴, while daily satellite observations^{5,6} and continuous *in situ* measurements^{7,8} are needed to observe the emission and transport of dense aerosol plumes downwind of populated and polluted regions (urban haze), regions with vegetation fires (smoke), and deserts (dust). The effect of greenhouse gases on the energy budget occurs everywhere around the globe. Aerosols have both regional and global impacts on the energy budget, requiring frequent global measurements tied to elaborate models that provide realistic representations of the atmospheric aerosols^{3,9,10}.

Aerosol effects on climate differ from those of greenhouse gases in two additional ways. Because most aerosols are highly reflective, they raise our planet's albedo, thereby cooling the surface and effectively offsetting greenhouse gas warming by anywhere from 25 to 50% (refs 1, 9–11). However, aerosols containing black graphitic and tarry carbon

particles (present in smoke and urban haze) are dark and therefore strongly absorb incoming sunlight. The effects of this type of aerosol are twofold, both warming the atmosphere and cooling the surface before a redistribution of the energy occurs in the column. During periods of heavy aerosol concentrations over the Indian Ocean¹² and Amazon Basin¹³, for example, measurements revealed that the black carbon aerosol warmed the lowest 2–4 km of the atmosphere while reducing by 15% the amount of sunlight reaching the surface. Heating the atmosphere and cooling the surface below reduces the atmosphere's vertical temperature gradient and therefore is expected to cause a decline in evaporation and cloud formation^{14,15}.

The second way in which aerosols differ from greenhouse gases is through the aerosol effect on clouds and precipitation. In polluted regions, the numerous aerosol particles share the condensed water during cloud formation, therefore reducing cloud droplet size by 20–30%, causing an increase in cloud reflectance of sunlight by up to 25% (refs. 2, 16–19), and cooling the Earth's surface. The smaller, polluted cloud droplets are inefficient in producing precipitation^{20,21}, so they may ultimately modify precipitation patterns in populated regions that are adapted to present precipitation rates. The cooling effect due to polluted clouds is still poorly characterized with an uncertainty 5 to 10 times larger than the uncertainty in the predicted warming effect of greenhouse gases^{1,22}. The effect of aerosols on precipitation is even less well understood.

To assess the aerosol effect on climate we first need to distinguish natural from anthropogenic aerosols. Satellite data and aerosol transport models show that plumes of smoke and regional pollution have distinguishably large concentrations of aerosols in particular of fine (submicron) size. In contrast, natural aerosol layers may have concentrated coarse dust particles and only widespread fine aerosols from oceanic and continental sources²³. The ability of satellites to observe the spatial distribution of aerosols^{24–26}, and to distinguish fine from coarse particles, can be exploited to separate natural from anthropogenic aerosols. *In situ* measurements of aerosol composition^{29,30} and size, models that assimilate

the measurements and information on population density and economic activities are needed to further quantify the anthropogenic aerosol component, and to relate it to specific sources.

Aerosol research is in transition from an exploratory phase to a global quantitative phase. The exploratory phase is dominated by the discovery of new aerosol-related processes. For example, the large concentration of black carbon emitted from vegetation fires and found in regional pollution in the tropics^{31–33} and its effect on slowing down the hydrologic cycle^{2,34}, or the effect of aerosol on reducing precipitation efficiency^{20,21,35,36} and counteracting regionally the greenhouse warming^{9,10,17}. In this phase, models are used to assess the potential of aerosol processes to affect the global climate^{10,37}. Because aerosols vary widely from region to region, a multiple-measurement approach is necessary to assess their impacts on global climate. Specifically, we require the use of long-term, detailed global measurements from satellites^{17,23,38–40}, distributed networks of ground-based instruments^{29,41,42}, and comprehensive regional experiments in clean⁴³ and polluted^{31,32} environments, that feed global aerosol and climate models^{14,23,44}.

Regional variability of aerosols

Most aerosols are regional in nature owing to their short lifetime, the regional distribution of sources, and the variability in their properties. Seasonal meteorological conditions determine how far aerosols are transported from their sources as well as how distributed they are vertically through the atmosphere. Elevated aerosol layers can be picked up by strong winds and transported from Africa or Asia to America and from America to Europe^{7,8}. Aerosol properties are modified during the transport by dry or wet deposition, in-cloud processes, and atmospheric chemical reactions.

Aerosol optical thickness (AOT) describes attenuation of sunlight by a column of aerosol, and thus serves as a measure of aerosol column concentration. In Fig. 1a,b, AOT is shown separately for fine and coarse aerosol for September 2000, using satellite techniques

discussed in Box 1; several typical regional aerosols are distinguished (see also Table 1).

Urban and industrial regional pollution (urban haze)

These mainly fine hygroscopic particles are found downwind of populated regions⁵ (regions a, c and e in Fig. 1a) in air polluted, for example, by car engines, industry, cooking and fireplaces. In China, economic growth and population expansion increased AOT from 0.38 in 1960 to 0.47 in 1990 (ref. 45). Pollution aerosol was modelled first as sulphates only¹¹, but new chemical measurements⁴⁶ show that downwind of the eastern United States⁴⁷ the contribution of carbonaceous material to AOT (30%) is double that of sulphates (16%), with water intake (48%) and black carbon (6%) accounting for the rest. Black carbon describes the effective fraction of elemental carbon that accounts for the absorption properties of the aerosol. Emission of black carbon is lower for newer engine technology so that black carbon contributes generally more to AOT in south and east Asia and Central America^{32,33} (11%). Absorption by black carbon is not only related to its concentration, but also depends on its location in the aerosol particle—absorption can be two to three times stronger if the black carbon is located inside the scattering particle^{48–51}.

Smoke from vegetation fires

Smoke from vegetation fires is dominated by fine organic particles with varying concentrations of light-absorbing black carbon (regions b and d in Fig. 1a) emitted in the hot, flaming stage of the fire. In forest fires the flaming stage is followed by a long, cooler smouldering stage in which the thicker wood, not completely consumed, emits smoke (composed of organic particles without black carbon) in much greater quantities than during the flaming stage. Conversely, thin African grasses burn quickly in strong flaming fires, emitting large quantities of black carbon, without a smouldering stage. On average, 12% of African smoke AOT is due to absorption by black

Table 1 Climatology of ambient aerosol properties averaged on the atmospheric column

		Regional pollution aerosol				Biomass burning			Dust	Oceanic
Analysis\Aerosol type										
AERONET analysis		East. US	Europe	SE Asia	Cen. Am.	Boreal forest	Trop. forest	Savanna Africa–S.A.	Sahara–Saudi Arabia	Pacific Ocean
Time of the year		Jun–Sep	Jan–Apr	Jan–Apr	Jan–Dec	Jun–Nov			Jan–Dec	Jan–Dec
Average AOT		0.20	0.20	0.20	0.30	0.25–0.45	0.25–0.5		0.2–0.4	0.06
AOT _f		94%	95%	95%	90%	95%	92%		25%	67%
% absorption of AOT		3%	6%	12%		7%	12%		5%	2%
MODIS analysis		North Atlantic 60–105° W 20–45° N		SE Asia 70–140° E 5–40° N		South Africa 15° W–30° E 0–20° S			West Africa 15–50° W 10–25° N	
Average AOT		0.18		0.24		0.31			0.30	
AOT _f		41%		44%		66%			33%	
ΔF _{TOA} (W m ^{–2})		–8		–10		–10			–17	
ΔF _{SUR} (W m ^{–2})		–10		–23		–30			–23	

The aerosol properties presented are based on two types of analysis. The top part of the table shows systematic multi-year measurements by the Aerosol Robotic Network (AERONET)^{53,43} and *in situ* measurements^{47,57,58}, whereas data in the bottom part of the table are based on analysis of Moderate-resolution Imaging Spectroradiometer (MODIS) satellite data for September 2000. In the AERONET analysis, aerosol optical thickness (AOT) is a measure of the aerosol column concentration and is given at a wavelength of 0.55 μm. Four aerosol types are shown with a representative size distribution (in units of μm³ per μm²): (1) pollution from eastern United States (Greenbelt, Maryland) and Europe (Venice and Paris), southeast Asia (Maldives–INDOEX) and Central America (Mexico City); (2) biomass burning from Africa and South America; (3) dust over the Atlantic Ocean (Cape Verde); and (4) maritime aerosol over the Pacific Ocean (Lana). The uncertainty in AOT is ± 0.01; contribution of the fine mode to AOT (AOT_f) is given over the land with uncertainty of ± 2%. In the MODIS data the uncertainty in AOT ranges from ± 0.03 to ± 0.06. The reflection of sunlight to space (ΔF_{TOA}) and reduction of surface illumination (ΔF_{SUR}) are based on the AERONET and MODIS data, using the radiative transfer code of Chou⁵⁹.

carbon⁵², in contrast to 5% for Boreal fires³³. Smoke is less hygroscopic than regional pollution aerosol, with only 10–20% of smoke AOT being associated with water for typical relative humidities of 80–85% in South America and Africa³³, compared with 50% for regional pollution aerosol¹. Dense smoke plumes are found annually downwind of fires in South America (August–October), Central America

(April and May), Southern Africa (July–September) and Central Africa (January–March)^{5,6,54}.

Dust

Dust is emitted from dry lakebeds in the Sahara, east Asia and the Saudi Arabian deserts^{5,6} (regions a and c in Fig. 1b) that were flooded

Box 1

How do satellites observe aerosols?

Radiant energy reflected and emitted by the Earth carries with it a signature of the atmospheric and surface properties. By measuring the wavelength, angular and polarization properties of this energy, satellite sensors can quantify several atmospheric and surface properties. (Polarization is the degree of organization of the electric field of the scattered solar radiation into a given direction.)

The human eye is sensitive to a narrow spectral range of the solar spectrum, with spectral receptors for blue, green and red light. Depth perception is provided by our eyes having slightly different angles of observation. By analogy, aerosol remote sensing — for example, AVHRR (Advanced Very High Resolution Radiometer), METEOSAT and GOES (Geostationary Operational Environmental Satellite) — developed from using a single wavelength^{5,91} and single angle of observation, like a colour-blind person with one eye. The TOMS instruments (for Total Ozone Mapping Spectrometer), flown since 1978, have two channels sensitive to ultraviolet light that were discovered to be excellent for observations^{92,93} of elevated smoke or dust layers above scattering atmosphere (Box Fig. 1).

The first instrument designed for aerosol measurements, POLDER (Polarization and Directionality of the Earth's Reflectances)^{26,38,40}, uses a combination of spectral channels in a range wider than that of human vision (0.44–0.86 μm). The instrument comprises a wide-angle camera that observes the same target on the Earth at several different angles, up to zenith angle of 65°. POLDER also measures light polarization to detect fine aerosols over the land, taking advantage of the difference between the spectrally neutral polarized light reflected from the Earth's surface and the spectrally decreasing polarized light reflected by fine aerosols (see Fig. 4).

Two instruments, MODIS (Moderate-resolution Imaging Spectroradiometer) and MISR (Multi-angle Imaging Spectroradiometer), on the Terra satellite measure global aerosol concentrations and properties since 2000. Over the ocean, MODIS uses the aerosol spectral signature in a wide range (0.47–2.1 μm) to distinguish small pollution particles from coarse sea-salt and dust²⁵ (Box Fig. 2). MODIS measures aerosol optical thickness (AOT) with an estimated error of $\pm 0.05 \pm 0.20\text{AOT}$ over the land⁹⁴ and $\pm 0.03 \pm 0.05\text{AOT}$ over the ocean⁹⁵ (note that 0.03 is an offset due to uncertainty in the ocean state and 0.05AOT is due to uncertainty in aerosol properties). Over land, MODIS uses the 2.1- μm channel to observe surface-cover properties, estimate surface reflectance at

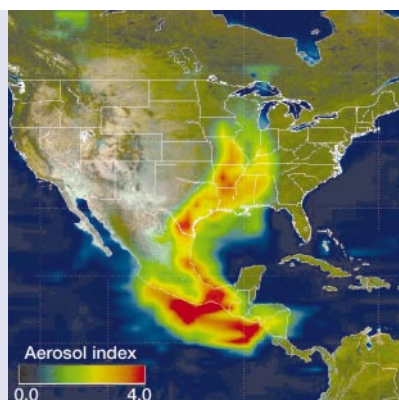
visible wavelengths, and derive AOT⁹⁶ from the residual reflectance at the top of the atmosphere.

The amount of light escaping the top of the atmosphere is affected by the angle at which the light was reflected by the surface or atmosphere. MISR²⁷ takes advantage of this fact by detecting the reflected light at different viewing angles (nadir to 70° forward and backward) along the satellite's track in a narrower spectral range (0.44–0.86 μm). It is thus able to separate the aerosol signal from that of surface reflectance, and determine the aerosol properties. A mixed approach using two viewing directions but a wider spectral range (0.55–1.65 μm) is used by ATSR (Along Track Scanning Radiometer)²⁸ to derive the aerosol concentration and type.

Over bright desert the magnitude of dust absorption determines if dust brightens or darkens the image^{82,97}, and this property is used for its estimation. Such satellite measurements, in agreement with *in situ*⁸, aircraft⁹⁸ and radiation-network³³ measurements of dust absorption, helped resolve a long-standing uncertainty in desert-dust absorption of sunlight⁵⁶.

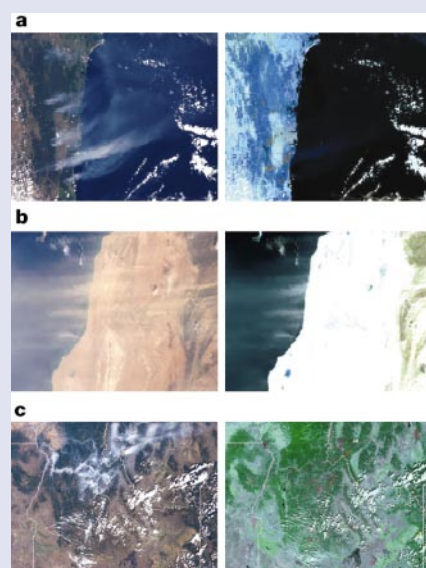
The realization that aerosols affect surface temperature and precipitation patterns creates a demand for more informative spaceborne observations. We foresee several improvements in this respect. First, spaceborne instruments are being considered that combine a wide spectral and viewing-angle range⁹⁹ with polarization. Such instruments are expected to improve derivation of fine and coarse aerosols, including their sizes and refractive index (refractive index, an optical property, is sensitive to the aerosol composition and water intake¹⁰⁰). Second, derivation of aerosol absorption over the oceans will be achieved by measuring the aerosol spectral attenuation of the bright glint.

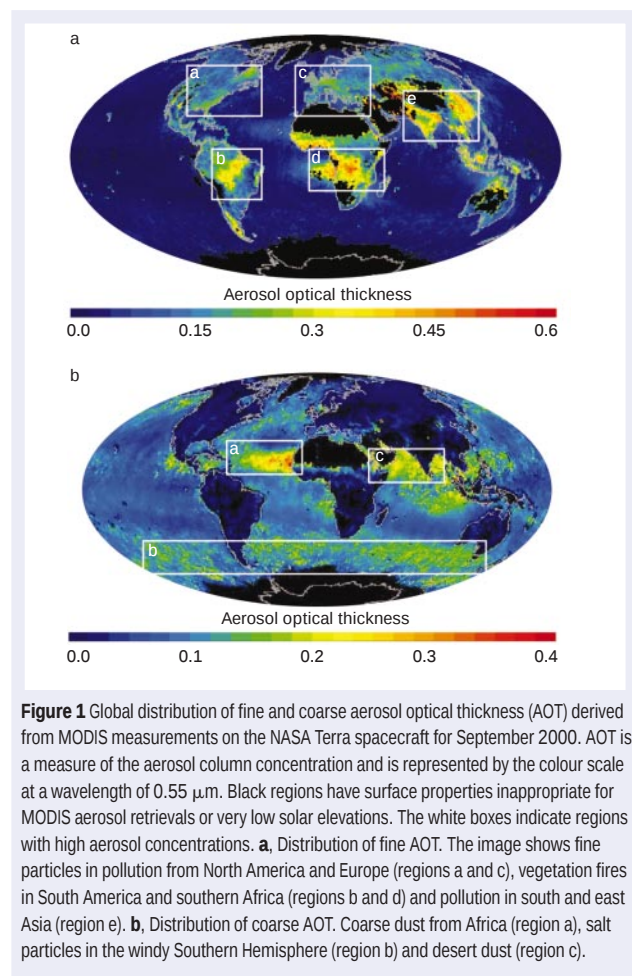
Box Figure 1 TOMS image showing heavy smoke aerosol transported to North America from large wild fires in Mexico on 15 May 1998. (Image courtesy of J. Herman, NASA/GSFC; see <http://toms.gsfc.nasa.gov/aerosols/aerosols.html> for additional images.)



Box Figure 2 Spectral aerosol reflectance measured by MODIS. Panels in the left column show colour composites of channels in the visible spectrum and those on the right show composites in the near infrared spectrum.

a, Fine smoke particles from fires in Australia (25 December 2001), which are invisible over the ocean in the near infrared²⁴. **b**, Coarse dust particles emitted from West Africa (7 January 2002), which are visible in both panels over the ocean. **c**, Smoke in Idaho and Montana invisible in the mid-infrared over the land. The spectral measurements are used to detect smoke over land^{31,95,96} and distinguish smoke from dust over the ocean.





in the Pleistocene era⁵⁵. Almost no dust is observed from Australia⁵, where the topography is mostly flat, because the arid regions are old and highly weathered⁵⁵. An unknown amount of dust is emitted from disturbed soils in Africa and east Asia. A previous estimate that 30–50% (ref. 56) of African dust results from human impact is questioned by new satellite observations using the ultraviolet part of the solar spectrum⁶ (Box 1); these show that African dust originates mostly from uninhabited regions north of 15° N (ref. 55). Dust emission from Africa is influenced by large-scale air circulation⁵⁷, which affects flow from the continent, and drought conditions. The highest dust production matches drought conditions in the strongest El Niño year of 1983 (ref. 7).

Dust AOT is dominated by coarse particles⁵⁸ with varying concentrations of iron oxide (rust) that absorbs light in the blue and ultraviolet wavelengths. However, African dust transported to Florida contains high concentrations of fine particles (10–100 $\mu\text{g m}^{-3}$) during the summer months and exceeds local pollution standards on particulate matter⁴². Dust from east Asia, from both natural sources and land use, is elevated to a height of 3–5 km with a pollution layer under it at 0–2 km (ref. 59), and is transported during April and May to North America. In April 2001 such a dust storm generated hazy conditions (AOT of 0.4) as far away as Boulder, Colorado (G. Feingold personal communication). On its way to North America, dust deposited in the Atlantic and Pacific Oceans provides key nutrients such as iron to oceanic phytoplankton⁶⁰.

Oceanic aerosol

Oceanic aerosol is shown in the blue regions of Fig. 1b and the elevated green values at latitudes south of 40° S (region b in Fig. 1b). It is

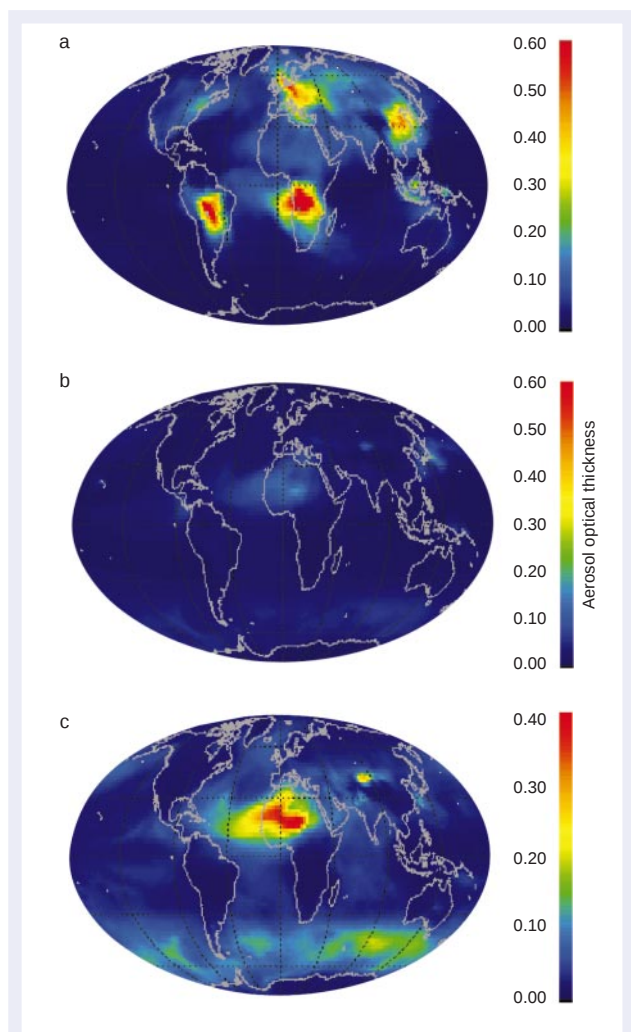


Figure 2 Model results of Chin *et al.*²³ that correspond to the MODIS data of September 2000. **a**, Anthropogenic fine aerosols; **b**, natural fine aerosols; and **c**, coarse aerosols composed of natural dust and salt.

composed of coarse salt particles emitted from bursting sea foam in windy conditions and fine sulphate particles from oceanic emissions⁶¹. Oceanic aerosol generally absorbs very little sunlight^{33,62} — its AOT is estimated to average 0.07 in most regions^{32,62}, but increases in the windy region south of 40° S to 0.2 (ref. 23).

Anthropogenic component

It is difficult to distinguish anthropogenic from natural aerosols as even individual particles can have both natural and anthropogenic components^{30,32,63}. Precise description of aerosol composition requires *in situ* chemical measurements that are restricted in time and location. However, it is possible to estimate the anthropogenic part of aerosols using a combination of satellite data, aerosol models^{23,24,44} and information on urban and agricultural activities and fire practices. For example, in Fig. 1 we see that anthropogenic aerosols downwind from vegetation fires and regions of industrial pollution are characterized by high concentration of fine particles. Aerosol models²³ confirm this finding, and further show that natural fine aerosols emitted from large-area oceanic and land sources exhibit much smaller spatial variability (Fig. 2). An example of daily satellite data and model calculations (Fig. 3) shows clear distinction between dust plume (coarse particles) and the regional pollution (fine particles). The model simulations confirm this distinction.

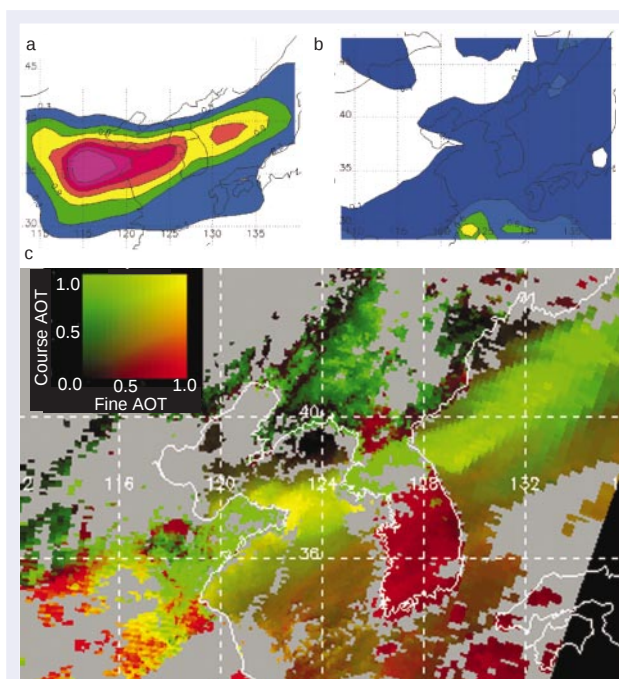


Figure 3 Satellite data and model calculations for a dust episode in east Asia advected over a pollution layer on 20 March 2001. **a, b**, NAAPS aerosol model simulation⁹⁰ of the optical thickness of the coarse dust particles (**a**) and the fine sulphate particles (**b**); the optical thickness varies from 0.3 for the blue to 1.5 for the red. **c**, Optical thickness of fine particles (red) associated with air pollution and the coarse dust (green) derived from the MODIS data on the Terra satellite.

An example of the relationship between population density and pollution concentration is shown in Fig. 4. Proximity of dust to desert or agricultural area can be used as an indicator for natural desert dust. Continued improvements in the models, constrained by new, detailed global measurements, will enable us to isolate the anthropogenic aerosol component.

In the Indian Ocean Experiment, chemical separation of aerosols into natural and anthropogenic components³² shows that the natural aerosol AOT is 0.07, with an additional 0.2–0.6 over the Bay of Bengal

resulting from anthropogenic activity⁶⁴. But pinpointing the anthropogenic source is proving difficult. Aerosol chemical composition can be used as a 'fingerprint' of the source, and a ratio of 1:2 between black carbon and total carbonaceous aerosol or sulphates suggests that the aerosol is emitted primarily from fossil fuel consumption³². But aerosol concentration varies through the year, whereas the relatively steady use of fossil fuel in the tropics, with its small variation of energy use from winter to summer⁶⁴, would suggest a stable aerosol concentration. A correlation between aerosol concentration and number of fires in India⁶⁴ suggests a large contribution of biomass burning. Analysis of aerosols and trace gases (CO and SO₂) suggests a mixed origin, both from fossil fuel and bio-fuel burning in the same proximity⁶⁵.

Aerosol direct forcing

The climate system varies naturally, through the dynamic interplay between atmospheric, oceanic and terrestrial moisture and energy. However, the radiative effects resulting from an increase in the concentration of anthropogenic aerosol or greenhouse gases, called radiative forcing, cause a net change in the Earth's absorbed and emitted solar and thermal energy and therefore are the basic ingredients of climate change¹.

A negative radiative forcing indicates that the Earth–atmosphere system loses radiant energy, resulting in cooling. Models of the climate system¹ show a direct relationship between radiative forcing and average global surface temperature, which rises 0.4–1.2 °C for every 1 W m⁻² of forcing. However, this relationship may break down for strongly absorbing aerosols^{12,14,15}.

Dust and smoke serve as an example of weakly and strongly absorbing aerosols (see Table 1). Absorption accounts for 5% and 12% of AOT for Saharan Desert dust and for aerosols in south and east Asia, respectively³³. In September 2000, dust reflected 17 W m⁻² of sunlight to space, thus reducing by an equivalent amount the energy available to heat the Earth–atmosphere system. At the same time, the dust reduced surface illumination by 23 W m⁻²; the difference (6 W m⁻²) is due to the absorption of sunlight. The aerosols in south and east Asia also reduced surface illumination by 23 W m⁻², but absorbed 13 W m⁻², reflecting only 10 W m⁻² to space³². Figure 5 shows the large regional extent of the aerosol radiative effects for these two aerosol types.

Another example of the radiative effects of weakly or strongly absorbing aerosol occurs over the Atlantic Ocean. Absorbing smoke is transported from Africa across the Southern Atlantic and less absorbing pollution aerosol is transported from North America in the Northern Atlantic (Fig. 1). In September 2000, aerosols found in these two

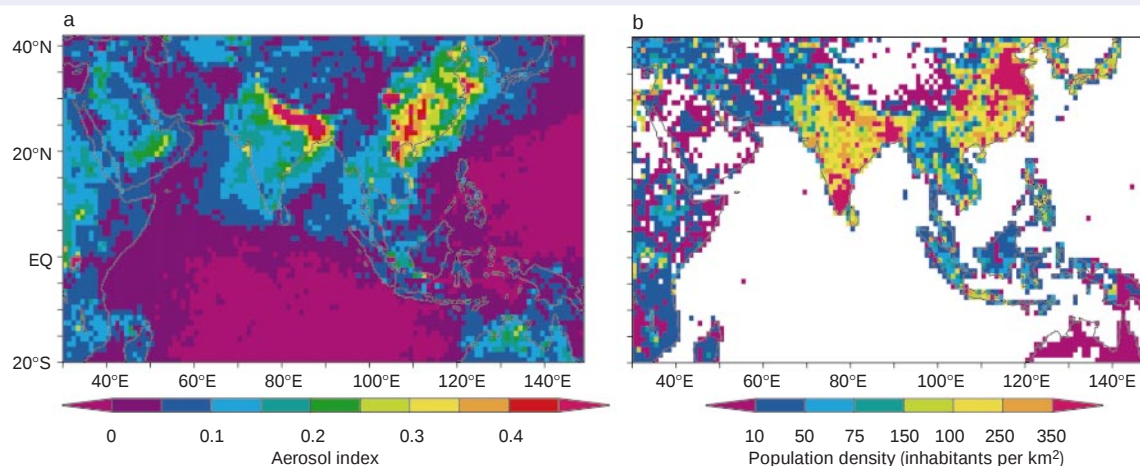


Figure 4 Comparison between concentration of anthropogenic aerosol and population density. **a**, Aerosol polarization index derived from the data of the POLDER instrument flown on ADEOS-1 in February 1997. The index is derived from measurements of

polarization of scattered solar light and is sensitive only to the presence of fine aerosol particles that in high concentration originate from anthropogenic sources. **b**, Population density map (inhabitants per square kilometre).

regions reflected 10 and 8 W m^{-2} , respectively, of sunlight to space. But the impact on surface illumination is three times larger for the smoke than for the pollution aerosol (30 W m^{-2} versus 10 W m^{-2}). Model calculations show that such strong aerosol absorption occurring in the lower troposphere heats the air layer, reducing its relative humidity and temperature gradients³², and increasing atmospheric stability. This decreases cloudiness¹⁵ and possibly reduces or prevents precipitation²⁰ that could have washed the aerosols from the atmosphere.

The results in Fig. 5 and Table 1 were computed for cloud-free oceanic regions. In cloudy conditions, the aerosol radiative effect depends on the fraction of absorbing aerosol located above clouds¹⁴, where the particles can absorb up to three times more sunlight than aerosol in cloud-free conditions⁶⁶. Absorbing aerosols may even have a positive (warming) rather than negative (cooling) radiative impact. Over dark oceans (albedo of 6%) the aerosol absorption has to be as high as 15% of AOT to result in warming, but 10% is enough over the brighter land (albedo of 20%), and only 5% if the aerosol layer is above boundary-layer clouds¹⁴.

Knowledge of the vertical distribution of aerosols and clouds is therefore needed to calculate the impact of aerosol radiative effects. Elevated aerosol plumes occur during long-range transport of aerosols. For example, smoke from forest fires in Canada was found over Greece at 2–3 km altitude⁶⁷, above the local boundary-layer clouds; Saharan dust is transported over Europe at 3–10 km elevation⁶⁸; and smoke intrusions from fires in Mexico⁶⁹ and northeast Canada are transported across North America at altitudes of 3–5 km. With the launch of space lidars⁷⁰ in the next few years we will begin to make the global measurements of the aerosol and cloud profiles needed to assess aerosol radiative forcing.

Aerosol modification of clouds and precipitation

Each cloud drop requires an aerosol particle to condense upon; clouds could not form otherwise. Thus the concentration, size and

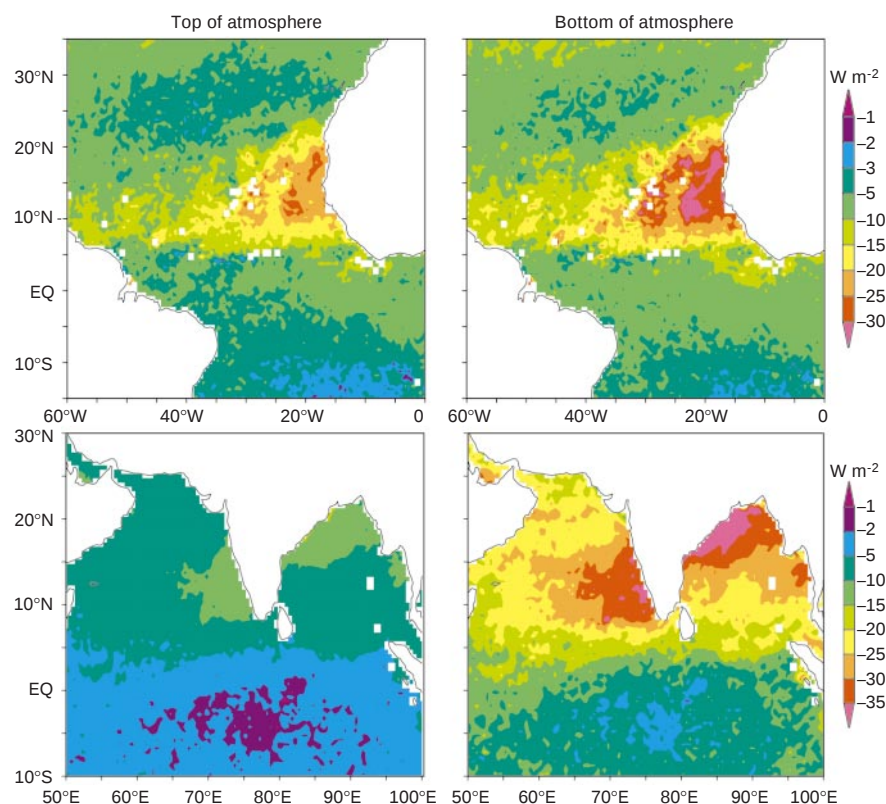
composition of the aerosols that can act as cloud condensation nuclei (CCN) determine the cloud properties^{9,17,18}, evolution and development of precipitation²⁰. However, availability of moisture, updrafts and cloud formation are influenced mainly by large-scale dynamic processes. Although natural aerosols are needed to form clouds, it seems that urban haze and smoke aerosols take every opportunity to reduce formation of precipitation and affect cloud radiative properties at the same time (Fig. 6).

Cloud base

Aircraft measurements show that in polluted air a sixfold increase in the number of fine aerosols per unit volume of air produces a three- to fivefold increase in the droplet concentration^{2,71}. Analysis of global satellite data shows that such a change in aerosol concentration corresponds to 10–25% smaller cloud droplets^{17,38} (Fig. 7), because the condensed water is divided into more numerous droplets. Clouds with smaller, more numerous droplets have a larger surface area and a corresponding increase in reflectivity of up to 30% (ref. 17). This increase in the reflection of sunlight to space, called the first aerosol indirect effect, was proposed by Twomey, based on two decades of aircraft sampling, to possibly rival the greenhouse forcing⁹. If these global cloud modifications can be attributed to anthropogenic effects⁷², they would translate into a solar radiative forcing of –0.5 to –1.5 W m^{-2} (refs 17, 73).

The actual effect of aerosols on cloud droplets, as inferred from global measurements from aircraft² and satellites^{17,38}, is 1.5–3 times smaller than that expected by Twomey for constant liquid water content⁹. In polluted air, above a given threshold of CCN concentration, the concentration of cloud droplets does not increase further; this results from competition between the more numerous CCN for water vapour², and it may explain the smaller global effect. The threshold depends on cloud dynamics, availability of moisture, aerosol size distribution and chemical properties. The larger and more hygroscopic the particles, the greater their ability to

Figure 5 Solar radiative perturbation at the top of the atmosphere and the surface for the tropical Atlantic and Indian Ocean. Top of the atmosphere (left parts) and surface (right parts) solar radiative perturbation (W m^{-2}) in clear sky is shown for the tropical Atlantic Ocean in May 1997 (upper parts) and the Indian Ocean in March 1997 (lower parts). The radiative perturbation is derived from the POLDER daily analysis of the aerosol optical depth combined with aerosol models modified to reproduce the aerosol absorption of Table 1. Over the Indian Ocean the perturbation is three times larger at the surface than at the top of atmosphere, as the aerosol-absorbed sunlight does not reach the surface.



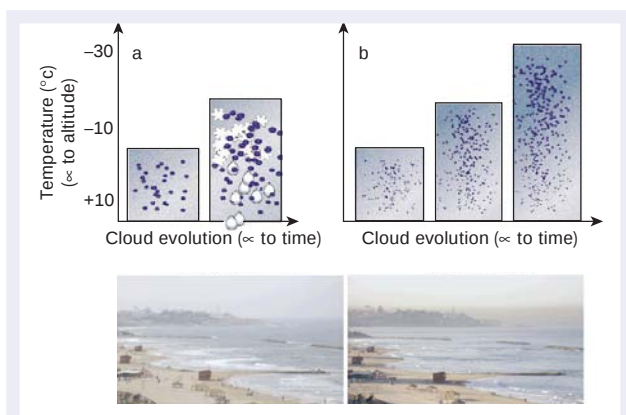


Figure 6 Schematic diagram of cloud formation in a clean and polluted atmosphere.

a, In a clean atmosphere, the cloud droplet size increases with cloud development until liquid precipitation or glaciation and precipitation take place. **b**, In polluted clouds, the availability of cloud condensation nuclei decreases cloud droplet development. In clouds with strong updrafts the developed cloud can be supercooled with no glaciation down to -37.5°C . The filled circles show the location of droplets of varying size, the asterisks show the location of ice crystals, and the oval shapes indicate rain drops.

compete at a given cloud updraft speed⁷⁴. The presence of even a few supermicron CCN particles per litre, such as sea salt or dust particles coated with sulphates⁶³, may reduce the indirect effect of pollution aerosol and produce precipitation⁷⁵.

Additional aerosol processes have to be considered, such as the possibility that water-soluble organic compounds present in the particle and the presence of soluble gases (HNO_3) in the atmosphere help the aerosols to take up water vapour and further increase the

concentration of cloud droplets and the indirect forcing⁷⁶. A contrary effect can take place from organic films that retard droplet growth and reduce drop concentration⁷⁷.

Cloud development

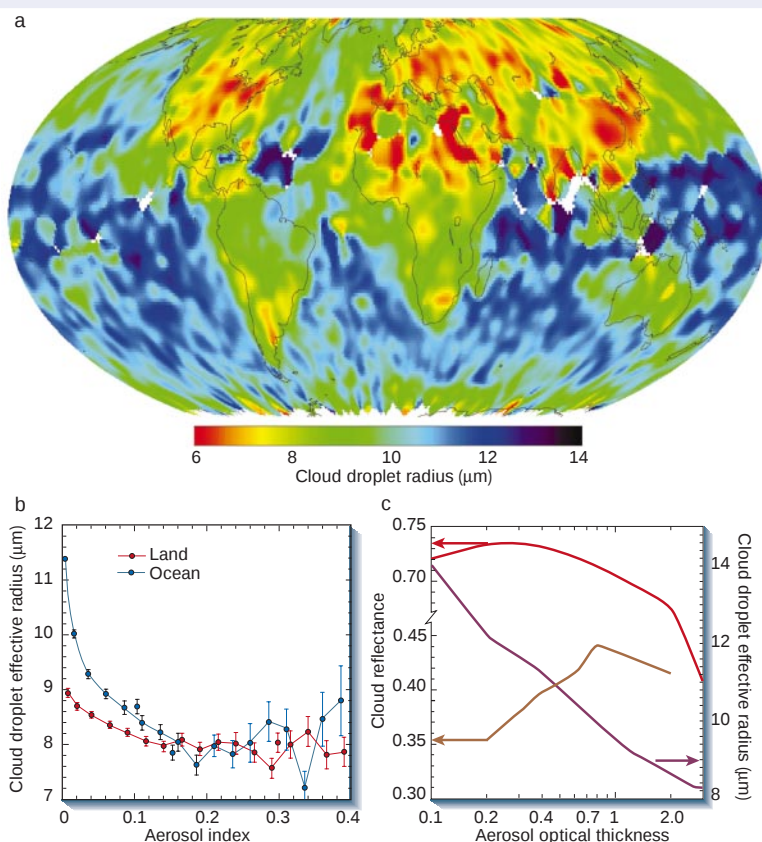
In clean conditions the cloud droplet size increases as the cloud develops and extends in the vertical direction until the droplet reaches a critical radius of $\sim 15\ \mu\text{m}$ (ref. 36) for the onset of liquid precipitation. Alternatively, the droplet may freeze if the cloud top temperature reaches -10°C . In pollution plumes over Australia and Canada, and smoke plumes over Indonesia, satellite data show not only smaller droplets at the cloud base ($5\text{--}8\ \mu\text{m}$ radius compared with $10\text{--}15\ \mu\text{m}$ in clean conditions), but also a lack of increase in droplet size as the cloud develops, rising through the atmosphere and accumulating water vapour. Consequently, precipitation does not occur or is delayed in polluted water clouds^{20,21}. In the same regions, non-polluted cloud droplets grow to $20\text{--}30\ \mu\text{m}$ and precipitation occurs. This suppression of precipitation was also observed for stratiform clouds polluted by emissions from ship stacks³⁶ and for polluted cumulus clouds in the Indian Ocean⁷¹.

The high concentration of aerosol supplies new CCN to condense the excess water vapour as the cloud cools down. The result is an increase in the cloud liquid water content, cloud lifetime and area of coverage — called the second aerosol indirect effect. The global importance of this effect is still not clear. Analysis of changes in cloud fraction and precipitation throughout the last century suggests that more clouds are needed for the same amount of precipitation, as would be expected given the inhibitory effect of pollution on precipitation (D. Rosenfeld personal communication, from data in ref. 1).

Ice crystals

Water clouds that cannot precipitate owing to the high concentration of aerosols could still precipitate once the droplets freeze.

Figure 7 Effect of aerosol on cloud droplet and reflectance derived from POLDER and AVHRR spaceborne measurements. **a**, Seasonal (March–May 1997) average droplet size in liquid water clouds estimated from the POLDER measurements³¹. **b**, The dependence of the droplet size on the aerosol index, also derived from POLDER over land (red) and ocean (blue). **c**, Analysis of AVHRR data for the dependence of the droplet size (purple) and cloud reflectance (brown and red) on aerosol optical thickness over the Amazon Basin during the dry burning season of 1987 (refs 16, 19). The reflectance of low-level clouds (brown) with reflectance of 0.35 increases with the aerosol concentration and the reflectance of bright clouds (red) decreases.



However, measurements in a deep convective cloud with a strong updraft show that the freezing process is postponed until the cloud reaches temperatures of -37.5°C . Strong updrafts and condensation lead to a high concentration of small cloud droplets. These droplets do not collide efficiently to form raindrops, nor do they freeze at -10°C . The result is a supercooled water cloud, thus eliminating an alternative way for clouds to precipitate⁷⁸. Model simulations suggest that this process requires the presence of high CCN concentrations ($1,260\text{ cm}^{-3}$) and vigorous updrafts⁷⁹.

Cloud formation

The presence of light-absorbing black carbon in aerosols can also affect cloud properties. Models show that heating of the lower troposphere by aerosol absorption reduces cloud formation¹⁵, an effect referred to as the semi-direct effect¹⁴. There are no direct measurements of this effect, but analysis of satellite data of clouds embedded in varying concentrations of smoke from fires in the Amazon basin¹⁹ shows that for the thicker clouds an increase in the smoke AOT from 0.2 to 3 raises the cloud-top temperature by 4°C , decreases the cloud reflectance by 0.13, while still reducing the droplet size by 40% (Fig. 7c). The simultaneous rise in the cloud-top temperature and reduction in reflectance — more than black carbon absorption itself can explain¹⁹ — indicates the possibility of a reduction in convection, thereby causing a decrease in the updraft speed and in the amount of liquid water available to form the cloud¹⁹.

The effect of aerosols on cloud droplet size is better understood than their effect on precipitation. Additional studies are needed to quantify the indirect effect of aerosols on climate, but the potential for a significant cooling in most regions is indicated. The reduction of the precipitation efficiency by anthropogenic aerosols has the potential to shift precipitation away from polluted regions³⁴. Because the continents are more polluted than the ocean, this can cause a loss of fresh water over the continents, and in particular around populated regions. However, long-term regional studies that can measure the significance of this effect are still not available.

Aerosols in climate change

The cooling influence of aerosols on climate, directly through the reflection of sunlight to space¹⁰ and indirectly through changes in cloud properties⁹, has been appreciated for over a decade, and has triggered a large number of observations, simulations and analyses. The effect of anthropogenic aerosols is not limited to cooling by sulphates. Instead, carbonaceous compounds that include light-absorbing black carbon can be an important warming agent, and the sign of the temperature change from aerosols⁸⁰ can vary depending on the aerosols' radiative properties^{81,82} and their distribution over the dark ocean and reflective land. The cooling of the Earth's surface from absorbing aerosols (compared with the top of the atmosphere) and consequential warming of the atmosphere causes a flattened vertical temperature profile in the troposphere, which is expected to slow the hydrologic cycle², reduce evaporation from the surface and reduce cloud formation^{14,15}. It has also been suggested that as aerosols tend to reduce cloud droplet sizes^{17,38}, and hence precipitation²⁰, rain and snow may be shifted from highly polluted populated areas to the more pristine oceanic regions³⁴. This raises the question of whether there has been or will be a change in the availability of fresh water due to aerosols.

Future research will need to unravel the magnitude of the aerosol effect on clouds and precipitation, on regional and global scales, and its sensitivity to aerosol chemistry and cloud dynamics. Given the importance of the absorption of sunlight by black carbon¹², a quantitative assessment of black carbon sources⁸³, its lifetime in the atmosphere, its distribution around the globe⁵⁴ and its impact on the hydrologic cycle² will need to be explored.

To associate the aerosol impact with human activity, we need to distinguish natural from anthropogenic aerosols. By measuring separately fine and coarse particles, remote sensors distinguish the

emission and transport of dust (mostly from natural sources) from pollution and smoke aerosols (mostly anthropogenic) around the planet. Remote sensors also map the distribution and properties of clouds^{17,24,38,72}, precipitation^{20,21} and the Earth's reflected solar and emitted thermal energy to space^{39,84,85}, as these atmospheric constituents are impacted by the aerosol. These global data and source characterization⁸³ feed aerosol models^{11,23,86} to show us an increasingly realistic picture of aerosols around the world and their impact on the environment. To achieve these ends, ground-based and *in situ* measurements, models and satellite observation have to be each improved and integrated. □

doi:10.1038/nature01091

- Intergovernmental Panel on Climate Change. *Climate Change 2001—The Scientific Basis* (contribution of working group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change) (Cambridge Univ. Press, Cambridge, 2001).
- Ramanathan, V. *et al.* Aerosols, climate, and the hydrological cycle. *Science* **294**, 2119–2124 (2001).
- Andreae, M. O. *et al.* External mixture of sea salt, silicates, and excess sulfate in marine aerosols. *Science* **232**, 1620–1623 (1986).
- Keeling, C. D. The concentration and isotopic abundances of carbon dioxide in the atmosphere. *Tellus* **12**, 200–203 (1960).
- Husar, R. B., Prospero, J. & Stowe, L. L. Characterization of tropospheric aerosols over the oceans with the NOAA AVHRR optical thickness operational product. *J. Geophys. Res.* **102**, 16889–16909 (1997).
- Herman, J. R. *et al.* Global distribution of UV-absorbing aerosol from Nimbus-7/TOMS data. *J. Geophys. Res.* **102**, 16911–16922 (1997).
- Prospero, J. M. & Nees, R. T. Impact of the North African drought and El Niño on mineral dust in the Barbados trade wind. *Nature* **320**, 735–738 (1986).
- Clarke, A. D. & Charlson, R. J. Radiative properties of the background aerosol: absorption component of extinction. *Science* **229**, 263–265 (1985).
- Twomey, S. A., Piepgrass, M. & Wolfe, T. L. An assessment of the impact of pollution on the global albedo. *Tellus* **36B**, 356–366 (1984).
- Charlson, R. J. *et al.* Climate forcing of anthropogenic aerosols. *Science* **255**, 423–430 (1992).
- Kiehl, J. T. & Briegleb, B. P. The relative roles of sulfate aerosols and greenhouse gases in climate forcing. *Science* **260**, 311–314 (1993).
- Satheesh, S. K. & Ramanathan, V. Large differences in tropical aerosol forcing at the top of the atmosphere and Earth's surface. *Nature* **405**, 60–63 (2000).
- Eck, T. F. *et al.* Measurements of irradiance attenuation and estimation of the aerosol single scattering albedo for biomass burning in Amazonia. *J. Geophys. Res.* **103**, 31865–31878 (1998).
- Hansen, J., Sato, M. & Ruedy, R. Radiative forcing and climate response. *J. Geophys. Res.* **102**, 6831–6864 (1997).
- Ackerman, A. S. *et al.* Reduction of tropical cloudiness by soot. *Science* **288**, 1042–1047 (2000).
- Kaufman, Y. J. & Fraser, R. S. Confirmation of the smoke particles effect on clouds and climate. *Science* **277**, 1636–1639 (1997).
- Nakajima, T. *et al.* A possible correlation between satellite-derived cloud and aerosol microphysical parameters. *Geophys. Res. Lett.* **28**, 1171–1174 (2001).
- Coakley, J. A. Jr, Bernstein, R. L. & Durkee, P. A. Effect of ship stack effluents on cloud reflectance. *Science* **237**, 953–956 (1987).
- Kaufman, Y. J. & Nakajima, T. Effect of Amazon smoke on cloud microphysics and albedo. *J. Appl. Meteorol.* **32**, 729–744 (1993).
- Rosenfeld, D. TRMM observed first direct evidence of smoke from forest fires inhibiting rainfall. *Geophys. Res. Lett.* **26**, 3105–3108 (1999).
- Rosenfeld, D. Suppression of rain and snow by urban and industrial air pollution. *Science* **287**, 1793–1796 (2000).
- Boucher, O. & Haywood, J. On summing the components of radiative forcing of climate change. *Clim. Dynam.* **18**, 297–302 (2001).
- Chin, M. *et al.* Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and Sun photometer measurements. *J. Atmos. Sci.* **59**, 461–483 (2002).
- King, M. D. *et al.* Remote sensing of tropospheric aerosols from space: past, present and future. *Bull. Am. Meteorol. Soc.* **80**, 2229–2259 (1999).
- Tanré, D. *et al.* Remote sensing of aerosol over oceans from EOS-MODIS. *J. Geophys. Res.* **102**, 16971–16988 (1997).
- Deuze, J.-L. *et al.* Estimate of the aerosol properties over the ocean with POLDER. *J. Geophys. Res.* **105**, 15329–15346 (2000).
- Diner, D. J. *et al.* MISR aerosol optical depth retrievals over southern Africa during the SAFARI-2000 dry season campaign. *Geophys. Res. Lett.* **28**, 3127–3130 (2001).
- Veefkind, J. P., de Leeuw, G. & Durkee, P. A. Retrieval of aerosol optical depth over land using two-angle view satellite radiometry during TARFOX. *Geophys. Res. Lett.* **25**, 3135–3138 (1998).
- Delene, D. I. & Ogren, J. A. Variability of aerosol optical properties at four North American surface monitoring sites. *J. Atmos. Sci.* **59**, 1135–1150 (2002).
- Artaxo, P. *et al.* Large scale aerosol source apportionment in Amazonia. *J. Geophys. Res.* **103**, 31837–31847 (1998).
- Kaufman, Y. J. *et al.* The Smoke, Clouds and Radiation experiment in Brazil (SCAR-B). *J. Geophys. Res.* **103**, 31783–31808 (1998).
- Ramanathan, V. *et al.* The Indian Ocean Experiment: an integrated analysis of the climate forcing and effects of the great Indo-Asian haze. *J. Geophys. Res.* **106**, 28371–28398 (2001).
- Dubovik, O. *et al.* Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *J. Atmos. Sci.* **59**, 590–608 (2002).
- Rotstayn, L. D., Ryan, B. F. & Penner, J. E. Precipitation changes in a GCM resulting from the indirect effects of anthropogenic aerosols. *Geophys. Res. Lett.* **27**, 3045–3048 (2000).
- Lohmann, U. & Feichter, J. Impact of sulfate aerosols on albedo and lifetime of clouds: a sensitivity study with the ECHAM4 GCM. *J. Geophys. Res.* **102**, 13685–13700 (1997).

36. Albrecht B. A. Aerosols, cloud microphysics, and fractional cloudiness. *Science* **245**, 1227–1230 (1989).
37. Penner, J. E., Dickinson, R. E. & O'Neill, C. A. Effects of aerosol from biomass burning on the global radiation budget. *Science* **256**, 1432–1433 (1992).
38. Bréon, F.-M., Tanré, D. & Generoso, S. Aerosol effect on cloud droplet size monitored from satellite. *Science* **295**, 834–838 (2002).
39. Wielicki, B. A. *et al.* Clouds and the earth's radiant energy system (CERES): an earth observing system experiment. *Bull. Am. Meteorol. Soc.* **77**, 853–868 (1996).
40. Herman, M. *et al.* Remote sensing of aerosol over land surfaces including polarization measurements and application to POLDER measurements. *J. Geophys. Res.* **102**, 17039–17049 (1997).
41. Holben, B. N. *et al.* An emerging ground based aerosol climatology: aerosol optical depth from AERONET. *J. Geophys. Res.* **106**, 12067–12097 (2001).
42. Prospero, J. M. Long term measurements of the transport of African mineral dust to the Southern US: implications for regional air quality. *J. Geophys. Res.* **104**, 15917–15927 (1999).
43. T. S. *et al.* International Global Atmospheric Chemistry (IGAC) project's first aerosol characterization experiment ACE-1: overview. *J. Geophys. Res.* **103**, 16297–16318 (1998).
44. Takemura, T. *et al.* Single-scattering albedo and radiative forcing of various aerosol species with a global three-dimensional model. *J. Clim.* **15**, 333–352 (2002).
45. Luo Y. F. *et al.* Characteristics of the spatial distribution and yearly variation of aerosol optical depth over China in last 30 years. *J. Geophys. Res.* **106**, 14501–14513 (2001).
46. Novakov, T., Hegg, D. A. & Hobbs, P. V. Airborne measurements of carbonaceous aerosols on the East Coast of the United States. *J. Geophys. Res.* **102**, 30023–30030 (1997).
47. Hegg, D. A. *et al.* Chemical apportionment of aerosol column optical depth off the mid-Atlantic coast of the United States. *J. Geophys. Res.* **102**, 25293–25303 (1997).
48. Ackerman, T. P. & Toon, O. B. Absorption of visible radiation in atmosphere containing mixtures of absorbing and non-absorbing particles. *Appl. Opt.* **20**, 3661–3668 (1981).
49. Martins, J. V. *et al.* Effects of black carbon content, particle size, and mixing on light absorption by aerosol particles from biomass burning in Brazil. *J. Geophys. Res.* **103**, 32041–32050 (1998).
50. Haywood, J. & Boucher, O. Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: a review. *Rev. Geophys.* **38**, 513–543 (2000).
51. Jacobson, M. Z. Strong radiative heating due to the mixing state of black carbon in atmospheric aerosols. *Nature* **409**, 695–697 (2001).
52. Eck, T. F. *et al.* Characterization of the optical properties of biomass burning aerosols in Zambia during the 1997 ZIBBEE field campaign. *J. Geophys. Res.* **106**, 3425–3448 (2001).
53. Ross, J. L., Hobbs, P. V. & Holben, B. Radiative characteristics of regional hazes dominated by smoke from biomass burning in Brazil: closure tests and direct radiative forcing. *J. Geophys. Res.* **103**, 31925–31941 (1998).
54. Liouss, C. *et al.* A Three-dimensional model study of carbonaceous aerosols. *J. Geophys. Res.* **101**, 19411–19432 (1996).
55. Prospero, J. M. *et al.* Environmental characterization of global sources of atmospheric soil dust identified with the NIMBUS-7 TOMS absorbing aerosol product. *Geophys. Res.* (in the press).
56. Sokolik, I. N. & Toon, O. B. Direct radiative forcing by anthropogenic airborne mineral aerosol. *Nature* **381**, 681–683 (1996).
57. Moulin, C. *et al.* Control of atmospheric export of dust from North Africa by the North Atlantic oscillation. *Nature* **387**, 691–694 (1997).
58. Tanré, D. *et al.* Climatology of dust aerosol size distribution and optical properties derived from remotely sensed data in the solar spectrum. *J. Geophys. Res.* **106**, 18205–18218 (2001).
59. Clarke, A. D. & Kapustin, V. N. A Pacific aerosol survey, part 1: a decade of data on particle formation, transport, evolution, and mixing in the troposphere. *J. Atmos. Sci.* **59**, 363–382 (2002).
60. Gao, Y. *et al.* Seasonal distribution of Aeolian iron fluxes to the global ocean. *Geophys. Res. Lett.* **28**, 29–33 (2001).
61. Hoppel, W. A. *et al.* Aerosol size distribution and optical properties found in the marine boundary layer over the Atlantic Ocean. *J. Geophys. Res.* **95**, 3659–3686 (1990).
62. Smirnov, A. *et al.* Optical properties of atmospheric aerosol in maritime environments. *J. Atmos. Sci.* **59**, 501–523 (2002).
63. Levin, Z., Ganor, E. & Gladstein, V. The effects of desert particles coated with sulfate on rain formation in the Eastern Mediterranean. *J. Appl. Meteorol.* **35**, 1511–1523 (1996).
64. Leon, J. F. *et al.* Large-scale advection of continental aerosols during INDOEX. *J. Geophys. Res.* **106**, 28427–28439 (2001).
65. Reiner T. *et al.* Chemical characterization of pollution layers over the tropical Indian Ocean: signatures of emissions from biomass and fossil fuel burning. *J. Geophys. Res.* **106**, 28497–28510 (2001).
66. Liao, H. & Seinfeld, J. H. Effect of clouds on direct aerosol radiative forcing of climate. *J. Geophys. Res.* **103**, 3781–3788 (1998).
67. Formenti, P. *et al.* Aerosol optical properties and large-scale transport of air masses: observations at a coastal and a semi-arid site in the eastern Mediterranean during summer 1998. *J. Geophys. Res.* **106**, 9807–9826 (2001).
68. Gobbi, G. P. *et al.* Altitude-resolved properties of a Saharan dust event over the Mediterranean. *Atmos. Environ.* **34**, 5119–5127 (2000).
69. Kreidenweis, S. M. *et al.* Smoke aerosol from biomass burning in Mexico: hygroscopic smoke optical model. *J. Geophys. Res.* **106**, 4831–4844 (2001).
70. Winker, D. M., Couch, R. H. & McCormick, M. P. An overview of LITE: NASA's lidar in-space technology experiment. *Proc. IEEE* **84**, 164–180 (1996).
71. Heymsfield A. J. & McFarquhar, G. M. Microphysics of INDOEX clean and polluted trade cumulus clouds. *J. Geophys. Res.* **106**, 28653–28674 (2001).
72. Han, Q., Rossow, W. B. & Laci, A. A. Near global survey of effective droplet radii in liquid water clouds using ISCCP data. *J. Clim.* **7**, 465–496 (1994).
73. Boucher, O. GCM estimate of the indirect aerosol forcing using satellite-retrieved cloud droplet effective radii. *J. Clim.* **8**, 1403–1409 (1995).
74. Feingold, G. *et al.* Analysis of smoke impact on clouds in Brazilian biomass burning regions: an extension of Twomey's approach. *J. Geophys. Res.* **106**, 22907–22922 (2001).
75. Feingold, G. *et al.* The impact of giant cloud condensation nuclei on drizzle formation in stratocumulus: implications for cloud radiative properties. *J. Atmos. Sci.* **56**, 4100–4117 (1999).
76. Charlson, R. J. *et al.* Reshaping the theory of cloud formation. *Science* **292**, 2025–2026 (2001).
77. Feingold, G. & Chuang, P. Y. Analysis of the influence of film-forming compounds on droplet growth: implications for cloud microphysical processes and climate. *J. Atmos. Sci.* **59**, 2006–2018 (2002).
78. Rosenfeld, D. & Woodley, W. L. K. Deep convective clouds with sustained supercooled liquid water down to -37.5°C . *Nature* **405**, 440–442 (2001).
79. Khain, A. P., Rosenfeld, D. & Pokrovsky, A. Simulating convective clouds with sustained supercooled liquid water down to -37.5°C using a spectral microphysics model. *Geophys. Res. Lett.* **28**, 3887–3890 (2001).
80. Alpert, P. *et al.* Quantification of dust-forced heating of the lower troposphere. *Nature* **395**, 367–370 (1998).
81. Claguin, T. *et al.* Uncertainties in assessing radiative forcing by mineral dust. *Tellus* **50B**, 491–505 (1998).
82. Kaufman, Y. J. *et al.* Satellite and ground-based radiometers reveal much lower dust absorption of sunlight than used in climate models. *Geophys. Res. Lett.* **28**, 1479–1483 (2001).
83. Bond, T. C., Charlson, R. J. & Heintzenberg, J. Quantifying the emission of light-absorbing particles: measurements tailored to climate studies. *Geophys. Res. Lett.* **25**, 337–340 (1998).
84. Christopher, S. A. *et al.* First estimates of the radiative forcing of aerosol generated from biomass burning using satellite data. *J. Geophys. Res.* **101**, 21265–21273 (1996).
85. Li, Z. & Kou, L. Atmospheric direct radiative forcing by smoke aerosols determined from satellite and surface measurements. *Tellus* **50B**, 543–554 (1998).
86. Collins, W. D. *et al.* Simulating aerosols using a chemical transport model with assimilation of satellite aerosol retrievals: methodology for INDOEX. *J. Geophys. Res.* **106**, 7313–7336 (2001).
87. Satheesh, S. K. *et al.* A model for the natural and anthropogenic aerosols over the tropical Indian Ocean derived from Indian Ocean Experiment data. *J. Geophys. Res.* **104**, 27421–27440 (1999).
88. Remer, L. A. *et al.* Urban/industrial aerosol: ground based sun/sky radiometer and airborne in situ measurements. *J. Geophys. Res.* **102**, 16849–16859 (1997).
89. Chou, M. D. A solar-radiation model for use in climate studies. *J. Atmos. Sci.* **49**, 762–772 (1992).
90. Liu M., & Westphal, D. L. A study of the sensitivity of simulated mineral dust production to model resolution. *J. Geophys. Res.* **106**, 18099–18112 (2001).
91. Fraser, R. S., Kaufman, Y. J. & Mahoney, R. L. Satellite measurements of aerosol mass and transport. *Atmos. Environ.* **18**, 2577–2584 (1984).
92. Hsu, N. C., Herman, J. R. & Weaver, C. Determination of radiative forcing of Saharan dust using combined TOMS and ERBE. *J. Geophys. Res.* **105**, 20649–20661 (2000).
93. Torres O. *et al.* A long-term record of aerosol optical depth from TOMS observations and comparison to AERONET measurements. *J. Atmos. Sci.* **59**, 398–413 (2002).
94. Chu, D. A. *et al.* Validation of MODIS aerosol optical depth retrieval over land. *Geophys. Res. Lett.* (in the press).
95. Remer, L. A. *et al.* Validation of MODIS aerosol retrieval over ocean. *Geophys. Res. Lett.* (in the press).
96. Kaufman, Y. J. *et al.* Remote sensing of tropospheric aerosol from EOS-MODIS over the land using dark targets and dynamic aerosol models. *J. Geophys. Res.* **102**, 17051–17067 (1997).
97. Fraser, R. S. & Kaufman, Y. J. The relative importance of aerosol scattering and absorption in remote sensing. *IEEE J. Geosci. Rem. Sens.* **GE-23**, 525–533 (1985).
98. Fouquart, Y. *et al.* Observations of Saharan aerosols: results of ECLATS field experiment. II: Broadband radiative characteristics of the aerosols and vertical radiative flux divergence. *J. Clim. Appl. Meteorol.* **25**, 28–37 (1986).
99. Mishchenko, M. I. & Travis, L. D. Satellite retrieval of aerosol properties over the ocean using polarization as well as intensity of reflected sunlight. *J. Geophys. Res.* **102**, 16989–17013 (1997).
100. Chowdhary, J., Cains, B. & Travis, L. D. Case studies of aerosol retrieval over the ocean from multiangle, multispectral photopolarimetric remote sensing data. *J. Atmos. Sci.* **59**, 383–397 (2002).

Acknowledgements

We thank F. M. Bréon, M. Chin, O. Dubovik, G. Feingold, P. Formenti, M. Herman, D. Herring, B. N. Holben, S. Mattoo, L. Remer and D. Rosenfeld for measurements and calculations used in this paper and for editorial comments. POLDER was a CNES/NASDA project; TOMS and MODIS are NASA projects.

Constraints on future changes in climate and the hydrologic cycle

Myles R. Allen* & William J. Ingram†

*Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK (e-mail: myles.allen@physics.ox.ac.uk)

†The Hadley Centre, Met Office, London Road, Bracknell RG12 2SZ, UK (e-mail: william.ingram@metoffice.com)

What can we say about changes in the hydrologic cycle on 50-year timescales when we cannot predict rainfall next week? Eventually, perhaps, a great deal: the overall climate response to increasing atmospheric concentrations of greenhouse gases may prove much simpler and more predictable than the chaos of short-term weather. Quantifying the diversity of possible responses is essential for any objective, probability-based climate forecast, and this task will require a new generation of climate modelling experiments, systematically exploring the range of model behaviour that is consistent with observations. It will be substantially harder to quantify the range of possible changes in the hydrologic cycle than in global-mean temperature, both because the observations are less complete and because the physical constraints are weaker.

Climate prediction is about quantifying risks and probabilities. Changes in rainfall distributions could have far more impact than the more-often-cited risk of global warming, but systematically quantifying the likelihood of such changes is only beginning. We cannot hope to predict the exact climate of 2050, still less the weather on a particular day, but we can assess the relative likelihood of different long-term trends given the constraints provided by observations and physical understanding available today.

Here we will explore the range of possible externally driven changes in the hydrologic cycle that may have taken place over the twentieth century or be in prospect for the twenty-first, emphasizing the question of what can be said about future climate^{1,2}. We will not attempt a comprehensive review of the literature, which is amply provided by the recent Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change (IPCC)^{3,4}. Instead, we review the present status regarding quantitative estimation of the distribution of likely precipitation changes, and their impact on other aspects of the atmosphere–ocean system, focussing on the evidence provided by the twentieth century and its implications for the next few decades. Evidence from past climates may provide additional constraints, particularly on the more distant future. The interpretation of such data and the challenge of disentangling causes and effects are discussed by Kump on pages 188–190 of this issue.

We will focus on the impact of anthropogenic carbon dioxide, both for reasons of space and because CO₂ is expected to dominate climate change in the coming century⁴. Other anthropogenic factors are probably having significant impacts on the hydrologic cycle, notably various forms of aerosol (see ref. 5, and the review in this issue by Kaufman *et al.*, pages 215–223). Changes are also anticipated^{6–10}, and may have already been detected^{11,12}, in how precipitation is distributed between low- and high-intensity events⁷. Regional details of the precipitation response to greenhouse warming are much less clear, as are hydrologic feedbacks on other aspects of the climate system, notably the strength of the oceanic thermohaline circulation (refs 13, 14, and see review in this issue by Rahmstorf, pages 207–214).

Physical arguments, such as those discussed in the preceding articles, indicate the qualitative changes in the hydrologic cycle expected to result from a variety of external influences. They do not, however, provide much more than order-of-magnitude constraints on the likely size of these changes under anthropogenic climate change (ref. 15, and see review in this issue by Pierrehumbert, pages 191–198). Ideally, forecasts of climate change should be constrained objectively by uncontroversial physical principles, possibly but not necessarily through simulation models, combined with actual climate observations. This approach may already be yielding results for predictions of global-mean temperature^{16–20}, but extending it to regional temperature change or changes in the hydrologic cycle presents new challenges which may require a whole new generation of climate modelling experiments²¹.

Constraints on global-mean temperature change

Anticipating difficulties in placing quantitative constraints on changes in the hydrologic cycle, we shall begin with a better constrained problem to show what can be done. Forecast global-mean temperatures are likely to be better constrained by recent climate observations than almost any other climate variable, both for physical reasons and because the historical temperature record is more accurate than for any other global variable. The constraint of global energy conservation means that the global-mean temperature response to an increase in CO₂ is controlled largely by three basic properties of the climate system: (1) the strength of atmospheric and surface feedbacks, which determine the so-called ‘climate sensitivity’ (equilibrium warming on doubling CO₂); (2) the effective heat capacity of the fraction of the oceans in contact with the atmosphere on short (sub- to inter-annual) timescales; and (3) how heat export to the ocean depths depends on recent changes at the surface.

In the absence of a sudden nonlinear climate change (in, say, ocean circulation) these properties of the climate may be expected to change slowly, if at all, in response to an imposed external forcing. Indeed, they are assumed to be constant in most simple and intermediate prognostic models of climate^{22,23}. Even the most complex climate models suggest that strong nonlinearities are unlikely to be important in

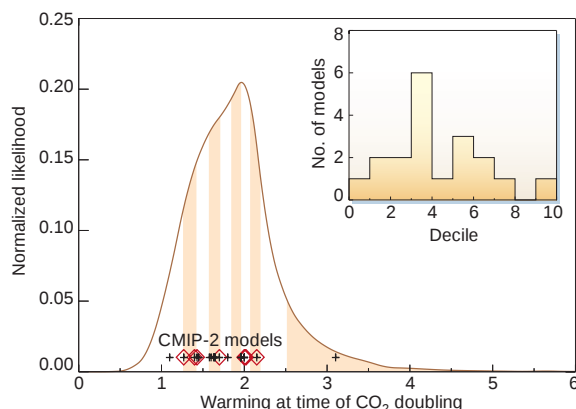


Figure 1 The range of transient climate response (TCR) consistent with recent observed temperature trends, compared to TCRs of some current AOGCMs. The curve shows the warming at the time of CO₂ doubling after an increase in CO₂ concentration of 1% per year (TCR), estimated from comparing an intermediate-complexity model with observations of recent large-scale temperature change, allowing for uncertainty due to internal variability as simulated by an AOGCM (refs 19, 31, with supplementary data supplied by M. D. Webster). The curve has been smoothed for clarity and the vertical bands show equal-area deciles of the distribution. The crosses are the TCRs of the AOGCMs in the CMIP-2 ensemble^{4,25}. Superimposed red diamonds show models used in the TAR summary range of '1.4–5.8 °C warming from 1990 to 2100'. The inset histogram shows how many of the CMIP-2 models fall into each decile of the observationally constrained distribution.

global-mean temperature change over the next half-century or so. For example, in all but one of the 19 coupled atmosphere–ocean general circulation models (AOGCMs) in CMIP-2 (the second Coupled Model Intercomparison Project^{24,25}), an annual 1% compound increase in CO₂ concentrations (a linear increase in radiative forcing) results in a near-linear global-mean temperature response up to and beyond the time of CO₂ doubling after a few years' initial adjustment (ref. 26; and see Fig. 9.3a of ref. 4). Likewise, global-mean precipitation increases roughly linearly in these experiments (Fig. 9.3b of ref. 4), albeit with rather stronger unforced and effective random variations from year to year.

The net radiative forcing in the CMIP-2 experiments is at the high end of projected anthropogenic changes over the coming decades. Hence current AOGCMs suggest that a strongly nonlinear global-mean temperature response to greenhouse forcing is unlikely over the next few decades at least. In that case, the constraint of global energy conservation means that estimates of past radiative forcing, recent observed near-surface temperature change and the accumulation of heat in the global oceans^{27,28} place objective, albeit still rather weak, constraints on the overall strength of atmospheric feedbacks^{16,17,19,20}. These in turn provide the basis for an objective probabilistic forecast of the temperature response to a given emissions scenario^{18,29,30} of the type we would like, ultimately, to provide for the hydrologic cycle.

The curve in Fig. 1 shows an estimate of the probability distribution of global-mean warming at the time of CO₂ doubling under a scenario of CO₂ concentration increasing by 1% annually (the 'transient climate response', or TCR), which is consistent with recent observations of large-scale surface, atmospheric and oceanic temperature change^{19,31}. Note that this empirical distribution is, if anything, likely to underestimate the range of uncertainty in TCR, as the analysis on which it was based assumed a negligible impact of natural forcing on temperature changes in the twentieth century¹⁷.

The crosses in Fig. 1 show the TCR of the 19 AOGCMs in the CMIP-2 multi-model comparison. If the CMIP-2 models were a

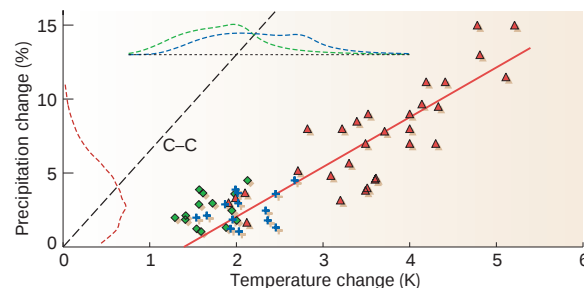


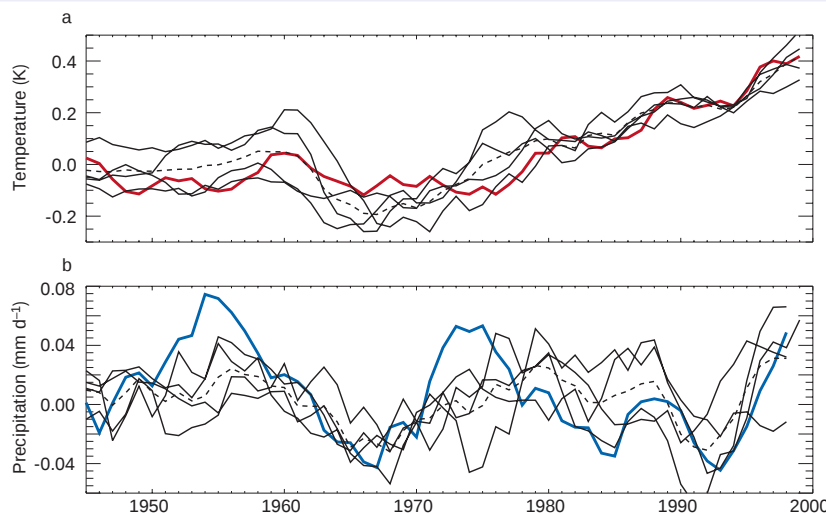
Figure 2 Global-mean temperature and precipitation changes in AOGCM simulations (scatter plots), and probability distributions obtained by requiring consistency with recent observations (curves). Red triangles show global-mean temperature and precipitation changes in a wide range of equilibrium CO₂-doubling experiments with simple thermodynamic ('slab') oceans^{4,45}, with the red line showing the best-fit (least squares) linear relationship. Green diamonds show the same, at the time of CO₂ doubling, for those CMIP-2 models for which the data are available²⁵. Blue crosses are the green diamonds adjusted for disequilibrium in the CMIP-2 runs by adding $\kappa F/k_T$ to ΔT (equation (2)), with a single value of κ ($=1$) estimated from the data to remove the bias with the best-fit line through the 'slab' experiments. All these points would lie on the dashed line labelled C–C if precipitation were to follow the Clausius–Clapeyron relation⁴⁴. The green dashed curve is the observationally constrained estimate of the distribution of global-mean temperature change at the time of CO₂ doubling from Fig. 1. The blue curve is the same, but adjusted for disequilibrium like the blue crosses. The red curve shows the distribution of global-mean precipitation changes implied by the blue curve, assuming the same straight-line relationship observed in the 'slab' experiments, with the same amount of scatter (assumed Gaussian).

random sample of possible climate-system behaviour consistent with these observations, then we should expect to find approximately equal numbers of models in each decile (vertical band) of the empirical TCR distribution and a more-or-less flat histogram in the inset panel. Instead, the models are concentrated near the centre of the distribution. Only one model displays a TCR in the uppermost two deciles of the distribution, and this turns out to be fortuitous. Warming accelerates in this particular model³² owing to some form of nonlinearity in the response. The empirical distribution (which assumes that both climate sensitivity and the nature of the ocean response are constant over these timescales) would immediately become much broader if it were to allow for such nonlinearity, pushing even this high-response model down into a relatively low percentile.

If current models underestimate the range of global-mean temperature responses consistent with recent observations, the problem can be expected to be worse for variables such as precipitation, which are not so well constrained by the available data. Hence any assessment of the risk of precipitation change exceeding a given threshold by a given date based solely on the spread of responses of currently available climate models¹⁰ will be underestimated, perhaps by a substantial margin.

Of course, the fact that current climate models do not span the range of responses consistent with recent warming is no indictment of the models: they were not designed to do so. The IPCC TAR was careful not to interpret the spread of the models as a direct measure of uncertainty in climate forecasts, for precisely this reason. Far from being designed to provide random samples of possible representations of the climate system, AOGCMs are generally designed as 'best guess' representations of the system based on a limited set of observations. Hence some clustering of model results towards the centre of the range of physically plausible behaviour should be expected. Because we cannot quantify the extent of this clustering *a priori*, we cannot predict the likelihood of the response in the real world lying above or below the range of model simulations with modelling alone. The only objective probabilistic forecast is provided by the

Figure 3 Changes in observed global-mean temperature (a) and land precipitation (b) over the past 55 years compared with a climate model. The model data are from a four-member initial-condition ensemble of the HadCM3 climate model run with estimates of anthropogenic, solar and volcanic forcing (dashed line shows ensemble mean). Model data were included only where and when observations exist and a five-year running mean was applied to suppress variability such as El Niño, which we would not expect to be coherent between them. Model simulations courtesy of P. Stott, The Met Office, and precipitation data courtesy of M. Hulme, Tyndall Centre^{38,71}.



constraint of consistency with actual climate observations. Can we identify analogous constraints on future changes in the hydrologic cycle?

Constraints on hydrologic indicators

Tropospheric humidity

The distribution of moisture in the troposphere (the part of the atmosphere that is strongly coupled to the surface) is complex, but there is one clear and strong control: moisture condenses out of supersaturated air. This constraint broadly accounts for the humidity of tropospheric air parcels above the boundary layer, because almost all such parcels will have reached saturation at some point in their recent history. Physically, therefore, it has long seemed plausible that the distribution of relative humidity would remain roughly constant under climate change³³, in which case the Clausius–Clapeyron relation implies that specific humidity would increase roughly exponentially with temperature. This reasoning is strongest at higher latitudes where air is usually closer to saturation, and where relative humidity is indeed roughly constant through the substantial temperature changes of the seasonal cycle. For lower latitudes it has been argued that the real-world response might be different (see ref. 34 and references therein). But relative humidity seems to change little at low latitudes under a global warming scenario, even in models of very high vertical resolution³⁵, suggesting this may be a robust ‘emergent constraint’ on which models have already converged.

Consistent with this picture of tropospheric specific humidities rising in line with surface warming and more-or-less unchanged relative humidity, apparently significant increases have been observed in surface^{36–38}, boundary-layer³⁹ and lower- to mid-tropospheric⁴⁰ specific humidity levels over some land regions in the Northern Hemisphere. In the tropics and Southern Hemisphere, *in situ* observations remain sparse and satellite records are too inhomogenous to estimate long-term humidity trends⁴¹. Estimated trends in upper tropospheric humidity — the most important for climate sensitivity, but also the most difficult to observe — are also ambiguous⁴². On shorter timescales, current climate-model simulations of the overall strength of water vapour feedback seem reasonably accurate in the case of the eruption of Mount Pinatubo, with the observed reduction in specific humidity being accompanied by only small changes in relative humidity⁴³.

Equilibrium global-mean precipitation changes

If tropospheric moisture loading is controlled by the constraints of (approximately) unchanged relative humidity and the

Clausius–Clapeyron relation, should we expect a corresponding exponential increase in global precipitation and the overall intensity of the hydrologic cycle as global temperatures rise? This is certainly not what is observed in models. The red triangles in Fig. 2 show the long-term equilibrium global-mean near-surface warming, ΔT , and precipitation increase, ΔP , in response to doubling CO_2 in a number of atmospheric models coupled to ‘slab’ (thermodynamic mixed layer) ocean models^{44,45}. They appear to lie, to a reasonable approximation, on a straight line⁴⁶ with slope 3.4% precipitation change per kelvin (much less than the 6.5% per kelvin implied by the Clausius–Clapeyron relation⁴⁴) and intersecting the temperature axis around 1.4 K. Why?

The explanation for these model results is that changes in the overall intensity of the hydrologic cycle are controlled not by the availability of moisture, but by the availability of energy^{44,47}: specifically, the ability of the troposphere to radiate away latent heat released by precipitation. At the simplest level, the energy budgets of the surface and troposphere can be summed up as a net radiative heating of the surface (from solar radiation, partly offset by radiative cooling) and a net radiative cooling of the troposphere to the surface and to space (R) being balanced by an upward latent heat flux (LP , where L is the latent heat of evaporation and P is global-mean precipitation): evaporation cools the surface and precipitation heats the troposphere.

We can approximate the perturbation energy budget of the troposphere as⁴⁷

$$\Delta R_c + \Delta R_t = L\Delta P \quad (1)$$

separating the perturbation radiative cooling, ΔR , into a component, ΔR_c , that is independent of ΔT and a component, ΔR_t , that depends on ΔT . The perturbation latent heating, $L\Delta P$, is about 1 W m^{-2} for a 1% increase in global precipitation. ΔR_c is the change in R that is due directly to external drivers of climate change (that is, change in R that is not mediated via the temperature response to these drivers, and hence approximately independent of ΔT). For example, doubling CO_2 decreases net outgoing infrared radiation through the tropopause (top of the troposphere) by $3\text{--}4 \text{ W m}^{-2}$ (depending on the details of stratospheric adjustment⁴), but also increases downward infrared flux at the surface by about 1 W m^{-2} (ref. 47), giving a negative ΔR_c of -2 to -3 W m^{-2} . In the absence of any significant change in tropospheric temperatures, therefore, increasing CO_2 concentrations reduce the intensity of the hydrologic cycle, an effect observed in early modelling experiments using prescribed sea surface

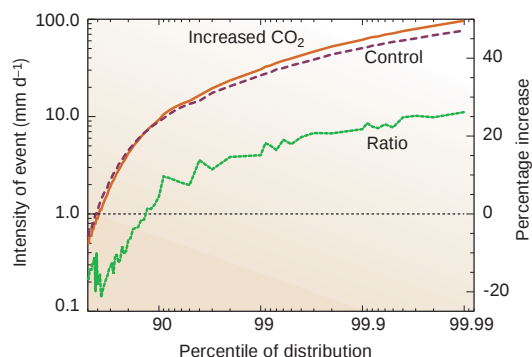


Figure 4 Log-log plot of the distribution of an AOGCM's daily precipitation and how it changes around the time of CO₂ doubling (J. M. Gregory, personal communication). Distribution of daily precipitation (all locations and seasons included) in years 2070–2100 of a transient climate-change simulation⁹⁵ (red) and in the corresponding control simulation (purple). Their ratio is shown by the green curve and right-hand (linear) axis. The global-mean warming is 3.6 K and the tropical mean 3.3 K, giving a Clausius–Clapeyron limit on this ratio of about 22%.

temperatures⁴⁸ (prescribing surface temperatures introduces an infinite heat sink at the surface, but the energy budget of the troposphere still has to balance).

The enhanced radiative cooling due to tropospheric warming, ΔR_T , is approximately proportional to ΔT : tropospheric temperatures scale with the surface temperature change and warmer air radiates more energy, so $\Delta R_T = k_T \Delta T$, with $k_T \approx 3 \text{ W m}^{-2} \text{ K}^{-1}$ (ref. 47). Changes in sensible heat flux at the surface are relatively small and the net effect of water vapour changes on k_T is small assuming (as discussed above) only small changes in the vertical profile of relative humidity⁴⁷. Contributions from clouds vary considerably between models³⁴ and presumably provide the main source of scatter in the relationship of ΔT versus ΔP shown in Fig. 2.

Both ΔR_C and ΔR_T are determined largely by basic features of the radiative and (in the case of ΔR_T) thermal response. Broad features of the pattern of warming, including a reduced equator–pole surface temperature gradient and reduced tropical ‘lapse rate’ (decrease of temperature with height) are much more consistent across climate models than the overall predicted magnitude of the change. Hence both ΔR_C and k_T are much less model-dependent and thus more reliable than is either ΔT or ΔP individually (another potentially useful emergent constraint). This simple picture can explain the main features of the equilibrium responses shown in Fig. 2, including a slope much less than that implied by the Clausius–Clapeyron relation and the line not intersecting the origin. It also has some interesting implications. The 10–90% range in equilibrium warming on doubling CO₂ that is consistent with recent large-scale warming is 1.8–6.5 K (ref. 19), based on the analysis behind Fig. 1. Translating this into a range in the equilibrium precipitation response using the best-fit line shown in Fig. 2 gives a 10–90% range of 0.6–18% (accounting for the scatter of points about the best-fit line, but ignoring uncertainty in the estimated slope, which would make the range slightly larger). Hence we cannot rule out, at even the 10% level, precipitation changes either above or below all currently available model predictions. As for the temperature response, the spread of available AOGCMs underestimates the only available objective estimates of uncertainty in future precipitation change.

Transient global-mean precipitation changes

A further consequence of equation (1) is that the expected precipitation change per degree of global warming, $\Delta P/\Delta T$, depends on the nature of the forcing. In a change driven by short-wave heating at the surface (for example, resulting from solar variability or scattering

aerosols), ΔR_C is small and $\Delta P/\Delta T$ is likely to be larger than in a CO₂-induced change. This effect is illustrated in Fig. 3 where observed global-mean temperature (Fig. 3a) and observed terrestrial precipitation (Fig. 3b) are compared over the past 55 years with the corresponding quantities diagnosed from an ensemble of four simulations using the Met Office's HadCM3 climate model forced with a combination of estimated anthropogenic and natural forcing⁴⁹.

Ensemble members and observations seem to move together, suggesting that both global-mean temperature⁴⁹ and, to a surprising extent, also continental-mean precipitation seem to be controlled by the external forcing. The correlation between the observed (5-year smoothed) precipitation time-series and the ensemble mean over this period is 0.55. This is greater than any correlation found in 98% of cases if we replace the observations with a similar-length, similarly sampled and smoothed segment of the HadCM3 control integration. Hence we can claim to have detected, in the simplest possible sense^{50–52}, the influence of external forcing on global-mean land precipitation.

It is, however, clear that terrestrial precipitation is not simply following the global temperature response. Precipitation changes seem to be dominated by the natural (solar and volcanic) forcing, which varies on shorter timescales, whereas the temperature response is dominated by the anthropogenic forcing, which increases comparatively steadily over this period. This is to be expected, as CO₂ is less effective in driving changes in global precipitation than is short-wave forcing of a similar magnitude because, in the former case, ΔR_C and ΔR_T in equation (1) tend to cancel each other out. Of course, both forcings could alter the distribution of precipitation between land and sea: we would prefer to compare global precipitation changes, but the necessary long-term data sets are not available³⁸.

Thus, although there is clearly usable information in Fig. 3, it would be physically unjustified to estimate $\Delta P/\Delta T$ directly from twentieth-century observations and assume that the same quantity will apply in the future, when the balance between climate drivers will be very different³³. Likewise, the approach of ref. 18 (estimating the size of the anthropogenic change from the observed record and using this to recalibrate anthropogenic changes simulated by climate models), although applicable in principle to precipitation, is unlikely to be useful at present. This is because the anthropogenic contribution to precipitation changes during the twentieth century is still weak compared with natural fluctuations (both unforced and externally driven). How, therefore, can we set about constraining the future transient response of the hydrologic cycle to anthropogenic forcing?

The transient responses of the CMIP-2 experiments, shown by the green diamonds in Fig. 2, seem to cluster above the distribution of equilibrium responses in the ΔT versus ΔP plane. Although the basic tropospheric energy budget displayed in equation (1) still applies, the relationship between ΔR_T and ΔT seems to be slightly different in the transient case. This discrepancy is probably due, at least in part, to systematic differences between transient and equilibrium patterns of warming: in particular, land areas tend to warm faster than oceans in transient experiments in all models. To first order, we might expect the discrepancy to be proportional to the net heat flux into the oceans in the transient experiments, F_s , which is a measure of the extent to which these experiments are out of equilibrium. Hence the tropospheric energy budget for a CMIP-2 scenario can be written

$$\Delta R_C + k_T \Delta T - \kappa F_s = L \Delta P \quad (2)$$

with ΔR_C and k_T taking values appropriate to the equilibrium response experiments and the empirical constant κ being estimated from the CMIP-2 results. The relationship between ΔP and ΔT in the CMIP-2 results is too weak either to confirm equation (2) or to constrain κ effectively, so we are simply proposing this as a plausible

Box 1

Towards objective probabilistic climate forecasting

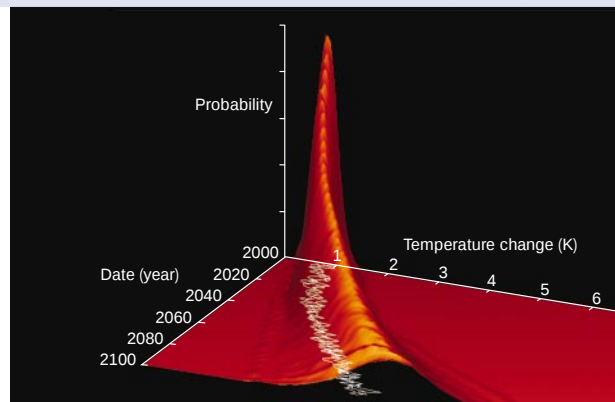
Myles R. Allen & David A. Stainforth

A climate forecast is intrinsically five-dimensional, spanning space, time and probability. Mainstream climate modelling has, so far, sacrificed probabilistic resolution almost completely in favour of the other four dimensions¹. Correcting this imbalance requires a new approach. Hitherto, forecasts have explored uncertainty in initial conditions^{2–3} (see Box 1 Figure opposite) as well as the impact of altering boundary conditions, such as adopting different scenarios of future concentrations of greenhouse gases⁴. For many variables, however, the main uncertainty in multi-decade climate prediction is not in the initial state nor in the external driving, but in the climate system's response⁵. Where this issue has been addressed with atmosphere–ocean general circulation models (AOGCMs) it is primarily through unconstrained 'ensembles of opportunity' based on comparisons between different models with no direct reference to observations^{4,6}. These, we have argued, are likely to provide a misleadingly small (over-optimistic) impression of forecast uncertainty⁷.

At the most basic level, the community needs to agree on what is meant by a 'correct' estimate of risk in climate forecasting. A probabilistic climate forecast (for example, of the risk of global precipitation increasing by more than 10% by 2050) cannot be checked against observations as can a probabilistic weather forecast. If a two-day weather forecast is wrong at the 5% level much more than 5% of the time, then something is amiss with the forecasting system. If a 50-year climate forecast is 'wrong' at the 5% level, we could simply have been unlucky. This apparent difficulty with verification has led to the view that estimates of uncertainty based on expert opinion (for example, the range for climate sensitivities used by ref. 8) are as good as any other⁹. This leaves the field open to the charge of subjectivism.

Objectively, we need to ask: 'given the observations available now, what range of forecasts might we have obtained had we started again from scratch many times over and made completely independent sets of decisions about model formulation and resolution?' A probabilistic forecast can only be said to have converged when including additional models is unlikely to make much difference to the forecast distribution of a particular variable. But there are complications. For example, we have no way of defining how 'close' two models are solely in terms of their formulation⁷, so we cannot design a representative sampling strategy over 'all possible AOGCMs' even if we had the resources to do so. In practice, therefore, a probabilistic forecast must begin with a very large ensemble of possible models, obtained by varying parameter values, parameterization schemes, resolution and entire model components, and extracting a sub-sample weighted according to the different models' ability to simulate recent observed climate change. A probabilistic forecast based on this sub-sample will have converged if its spread is determined primarily by the constraint of consistency with observations and not by the choice of models within the original ensemble — this is the crucial distinction between a constrained ensemble and an unconstrained ensemble of opportunity.

For some variables, probabilistic forecasts may be converging already. Given the emergent constraints relating past to future greenhouse warming that seem to hold across all available climate models, the distribution of forecast global-mean temperature changes in Fig. 1 is determined not by the choice of model(s), but by uncertainty in how much recent warming can be attributed to CO₂ increase. This uncertainty is due primarily to other signals and internal variability in the observed climate record. It will reduce as the signal strengthens⁵, but it may not change much as models improve. We would argue that this is both more robust (less subject to short-term revision) and more reliable (acceptable to non-specialists as a basis for action) than a forecast based on expert opinion.



Box 1 Figure The three-dimensional surface shows a forecast probability distribution of a one-dimensional quantity (global-mean warming above pre-industrial), accounting for uncertainty in the climate response, while the lines show the (smaller) impact of initial condition uncertainty in an ensemble of model simulations. Data courtesy of P. Stott (Met Office) and J. Kettleborough (Rutherford Appleton Laboratory)⁵, based on the IPCC SRES A2 scenario.

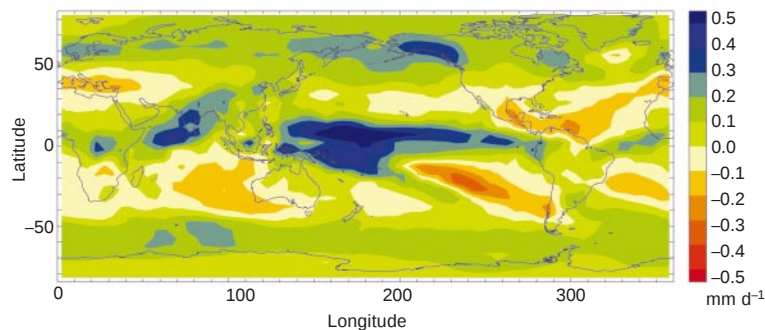
When will we be able to say that forecast changes in the hydrologic cycle have converged enough to be trusted? We made a tentative estimate of the distribution of global-mean precipitation change in Fig. 2, primarily to show that distributions based only on the spread of current AOGCMs should not be trusted. To extend this to regional changes, we need to repeat the analysis behind Fig. 1 with a full-scale AOGCM. Figure 1 required many hundreds of integrations to explore just three uncertain parameters in a two-dimensional climate model¹⁰, and there are hundreds of such uncertainties in an AOGCM. The chaotic nature of an AOGCM means that many of the techniques used in shorter-range forecasting to select perturbations¹¹ are not directly applicable to the climate problem¹²; it also means that several simulations will be needed to assess the impact of every perturbation to the model's formulation.

Thus, objective probabilistic forecasts of regional changes in rainfall and other climate variables will require numbers of simulations several orders of magnitude larger than the CMIP-2 experiment, the largest ensemble of AOGCM simulations undertaken to date. New approaches utilizing distributed computing and the emerging electronic 'grid' may provide a way forward^{13,14}, and readers interested in participating in such an initiative¹⁵ may wish to contact us on <http://www.climateprediction.net>.

Dave Stainforth is at Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK (e-mail: d.stainforth1@physics.ox.ac.uk)

- Schneider, S. H. *Nature* **411**, 17–19 (2001).
- Mitchell, J. F. B., Johns, T. C., Gregory, J. M. & Tett, S. F. B. *Nature* **376**, 501–504 (1995).
- Dai, A., Meehl, G. A., Washington, W. M., Wigley, T. M. L. & Arblaster, J. M. *Bull. Am. Meteorol. Soc.* **82**, 2377–2388 (2001).
- Cubasch, U. et al. in *Climate Change 2001, The Science of Climate Change* Ch. 9 (eds Houghton, J. T. et al.) 527–582 (Cambridge Univ. Press, Cambridge, 2001).
- Stott, P. A. & Kettleborough, J. A. *Nature* **416**, 723–726 (2002).
- Palmer, T. N. & Raisanen, J. *Nature* **415**, 512–514 (2002).
- Smith, L. A. *Proc. Natl Acad. Sci. USA* **99**, 2487–2492 (2002).
- Wigley, T. M. L. & Raper, S. C. B. *Science* **293**, 451–454 (2001).
- Giles, J. *Nature* **418**, 476–478 (2002).
- Forest, C. E. et al. *Science* **295**, 113–117 (2002).
- Palmer, T. N. *Rep. Prog. Phys.* **63**, 71–116 (2000).
- Lea, D. J., Allen, M. R. & Haine, T. W. N. *Tellus* **52A**, 523–532 (2000).
- Hansen, J. A., Allen, M. R., Stainforth, D. A., Heaps, A. & Stott, P. A. *World Resource Rev.* **13**, 187–189 (2001).
- Stainforth, D. A. et al. *Comput. Sci. Eng.* **4**, 82–89 (2002).
- Allen, M. R. *Nature* **401**, 642 (1999).

Figure 5 Precipitation change at the time of CO₂ doubling averaged over all the members of the CMIP-2 ensemble for which the data were available (14 models), for the 20 years centred on the time of CO₂ doubling.



correction for disequilibrium. The scatter of CMIP-2 results is consistent with what we would expect from the equilibrium response experiments. Thus, the problem is not too much noise, but too little signal — we do not have a wide enough range of ΔT or ΔP from the transient experiments to identify the relationship between them (see Box 1).

Drawing this evidence together, we can infer a distribution for ΔP at the time of CO₂ doubling under a scenario of an annual 1% increase in CO₂ concentration from the distribution of ΔT shown in Fig. 1 (the green dashed curve in Fig. 2) augmented with κF_s (blue dashed curve in Fig. 2, with the distribution of F_s estimated using the same analysis that yields Fig. 1; ref. 19). We translate this into a distribution for ΔP (red dashed curve in Fig. 2) using the best-fit line through the equilibrium response experiments, including additional variance to account for the scatter. It must be emphasized that these distributions use information on TCR and climate sensitivity only as constrained by observations of recent changes in temperature and ocean heat content. They do not depend on precipitation changes in the intermediate-complexity model used to derive the distribution function in Fig. 1 (which are likely to be model specific⁵⁴), nor do they make use of observed precipitation changes. It may seem counter-intuitive to use temperature and ocean heat content rather than precipitation observations to constrain future precipitation, but with global precipitation controlled by energetic constraints, and with the precipitation response to anthropogenic forcing being difficult to disentangle from other signals in the observations, this could be the best strategy available.

The precise distributions shown on Fig. 2 naturally depend on details of the analysis, but the key conclusion is that the current range of uncertainty in ΔP is considerably larger than the spread of current model responses. This analysis suggests a 25% chance of $\Delta P > 5\%$, the highest response of the coupled models shown in Fig. 2, and a 15% chance of $\Delta P > 6\%$, the response of the 'outlier' model mentioned above⁴ (for which the data needed for Fig. 2 are unavailable). The upper 5th percentile, often used in decision-making as a standard upper bound, could exceed the highest ΔP observed in current models by over 50%. The message is clear: impact assessments that rely on intermodel spread⁵⁵ could seriously underestimate the range of uncertainty in precipitation change in the twenty-first century.

A similar analysis could easily be extended to any variable that shows a simple functional relationship across a wide variety of models to quantities that are constrained by current observations, such as global temperatures and rates of ocean heat uptake over the coming decades. Of course there are caveats. In particular, the limited diversity of available slab and coupled AOGCMs makes it difficult to establish whether any relationship (such as the straight line in Fig. 2) has converged or is simply an artefact of resolution or sampling. Moreover, the use of an intermediate-complexity model to derive the forecast temperature distribution may underestimate the complexity of responses that are consistent with recent observations. Both of these caveats mean the spread of the forecast distribution of precipitation change is likely to be underestimated, so that the problems

with using inter-model spread would be even worse than suggested by the comparisons above. Both caveats could be substantially resolved with the use of much larger, systematically perturbed ensemble forecasts with AOGCMs (see Box 1).

Changes in precipitation extremes

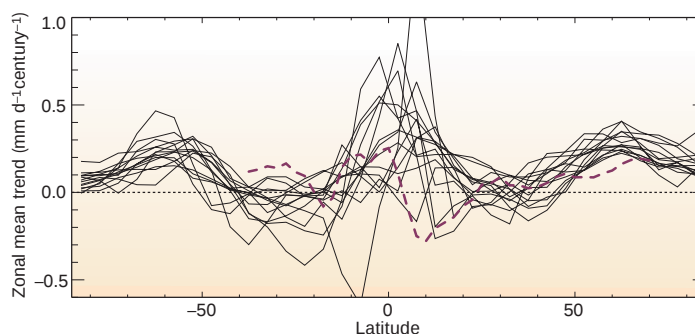
Long-term mean precipitation is a useful summary indicator of the intensity of the hydrologic cycle, but details of the distribution of precipitation over time, including the peak intensity of precipitation events and duration of prolonged droughts, are likely to be the most important issues in determining impacts of precipitation changes⁹. Although global-mean precipitation is primarily constrained by the energy budget, the heaviest rainfall events are likely to occur when effectively all the moisture in a volume of air is precipitated out, suggesting that the intensity of these events will increase with the availability of moisture⁷ (that is, significantly faster than the global mean). Thus we might expect the uppermost quantiles of the rainfall distribution to increase by about 6.5% per kelvin (ref. 44) if the ambient flows change (most likely at higher latitudes). In the tropics, where the flows leading to precipitation are themselves driven largely by the latent heat released by precipitation, larger increases still might occur⁷. In particular, the maximum thermodynamically possible rainfall and winds in hurricanes are predicted to increase rapidly with warming⁵⁶.

The red and purple curves in Fig. 4 show the magnitude of daily precipitation as a function of percentile of the precipitation distribution in one AOGCM under climate change and control conditions, respectively. The green curve (and right-hand axis) shows the ratio between them. At the highest end of the distribution (all tropical cases), it appears to be converging to about a 25% increase, which is indeed slightly more than we would expect from the Clausius–Clapeyron relationship.

Because such increases are more than double the increase in global-mean precipitation (that is, the change summed over all percentiles, which is constrained by the energy budget, not by the Clausius–Clapeyron relation), there must be decreases lower down the distribution. Indeed, the increase at the heaviest rain events is large enough that the energy constraint on the total implies that only on one day in ten does precipitation increase (the control and perturbed distributions cross around the 90th percentile in Fig. 4). It would be interesting to know how other models compare — if this were another emergent constraint generic over all models in regions of interest, then we could use forecast temperature changes to constrain extreme as well as mean precipitation. Given the acknowledged difficulties in relying on model simulations of extreme events and in observing changes in extremes directly⁵⁷, this could be a powerful result. In particular, this convergence to a particular fractional increase, possibly related to the Clausius–Clapeyron limit, could even improve as we move to the highest percentiles, which are generally the least tractable under more direct approaches.

Even modest increases in the magnitude of events in the tails of the distribution can have a very substantial impact on the expected

Figure 6 Zonal-mean precipitation trends over the 70 years to CO₂ doubling in the CMIP-2 ensemble and observed trends in terrestrial precipitation over the past century. Each thin solid line is for a member of the CMIP-2 ensemble (forced with CO₂ changes alone), whereas the thick dashed line is the observed trend over the twentieth century from land-based observations.



return times of events of a given magnitude^{8–10,58}. Such changes may already be occurring^{59,60}; the model used by ref. 61 suggested that, over the United Kingdom, floods of the magnitude expected to occur every 20 years in 1860 might now be expected every 5 years. Figure 4 implies that the corresponding changes might be even more significant in tropical regions¹⁰.

Regional precipitation

Although the large-scale temperature response to climate-change forcings is predicted to be relatively smooth (usually with the same sign everywhere), precipitation varies much more in space and time and is notoriously much harder to simulate correctly in models. It is no surprise then that predicted changes in local precipitation vary considerably both in space and between models. Figure 5 shows the CMIP-2 ensemble-mean pattern of regional precipitation change around the time of CO₂ doubling, while Fig. 6 shows the zonal-mean changes in each model.

The ensemble-mean precipitation change conceals a wide variety of responses across the CMIP-2 ensemble. Only 30% of the variance between models of predicted precipitation changes on a 5° × 5° grid is accounted for by a single common response pattern, compared to over 90% for temperature. The standard deviation between the various models' predictions is greater than the mean response shown in Fig. 5 everywhere between 50° N and 50° S apart from a small region in the equatorial West Pacific. The natural assumption that there can then be no useful information there may be over-pessimistic, as the small scales of precipitation variation tend to mask a closer agreement in physical terms. For example, if some set of models all shifted a particular rainband in the same direction, but its position varied between their base climates by as much as its width, one would see completely inconsistent results at a single point.

The zonal-mean response (Fig. 6) is more coherent across models, with strong increase in precipitation near the Equator, some reduction in sub-tropical subsidence regions and a smaller but more consistent increase in mid-latitudes. In short, the general tendency is for precipitation to increase in regions where it is high and decrease where evaporation is high, which is just what one might expect from the increased specific humidities leading to increased atmospheric fluxes of humidity from its sources to its sinks⁶². One consequence of this is that mid-continental summer droughts may become more frequent and intense⁶³, a common feature of model simulations of greenhouse warming^{64,65}. In interpreting any such features, however, it must be emphasized that imperfections in AOGCMs affect precipitation more than many other fields of general interest. For example, AOGCMs generally do not allow for the time convection takes to develop and hence simulate incorrectly the timing of precipitation over the course of a day. This can have knock-on effects on total precipitation.

The pattern of change shows a 'horseshoe' pattern reminiscent of the El Niño/Southern Oscillation (ENSO) phenomenon. Hence the interpretation of any trend is closely related to the question of whether the apparent⁶⁶, albeit debatable^{67,68}, increase in the frequency

and intensity of ENSO events over the past few decades is genuine and, if so, anthropogenic^{69,70}.

Figure 6 also shows the zonal-mean 100-year trend in observed precipitation from the Hulme dataset of land-based station observations^{38,71}. We see an increase in southern equatorial regions and in mid-latitudes, broadly consistent with the models and our physical expectations⁵³. But this is only a qualitative comparison, given that the land-based data sample less than a third of the world, the strongest precipitation changes in Fig. 5 seem to be occurring over the oceans, the forcing over the twentieth century is very different from the CMIP-2 experiments, and different forcings may well be associated with different patterns of response. Anthropogenic aerosols, in particular, are so geographically concentrated that circulation changes and other local effects might swamp the driving mechanism outlined above. Additionally, aerosols might affect precipitation more directly through their impact on cloud microphysics (ref. 5, and see review in this issue by Kaufman *et al.*, pages 215–223). Nevertheless, the apparent robustness of the basic features and agreement with a simple physical explanation are encouraging.

Because land-based precipitation records may miss most of the signal (Fig. 5), the advent of global satellite-derived precipitation analyses^{72–76} is welcome. Apparently significant positive trends in tropical precipitation and the strength of the Hadley circulation (the zonal-mean flow at low latitudes, fuelled by latent heat release) have been reported^{77,78}, although issues over decadal variability and the long-term stability of observationally based data sets remain^{79,80}.

An important factor for the regional details of the precipitation response to external forcing is the change in atmospheric circulation^{6,81}. In particular, an intensification of the North Atlantic Oscillation — more westerly winds across the North Atlantic and into Eurasia — has been observed over the past few decades^{81,82}. Although models (and simple theory) indicate that such a change, which would increase mid-latitude land precipitation, might be expected to accompany an anthropogenic greenhouse warming⁸³, most models seem to underestimate the magnitude of this circulation change⁸⁴. Extra-tropical circulation and precipitation anomalies may well be strongly influenced by driving from the tropics (refs 85, 86, and M. Blackburn and B. Hoskins, personal communication), and it remains very doubtful whether current AOGCMs respond correctly to localized but remote forcing^{83,87}.

Interactions with aerosols and oceans

We have concentrated above on the effects of CO₂, the largest single driver of temperature changes over the twentieth century and expected to be the dominant one in the twenty-first century⁸⁸. But aerosols also seem to have contributed significantly to recent large-scale temperature changes^{49,50} and may have had important effects locally. The most important overall are thought to be sulphate aerosols from volcanic and anthropogenic sources in the stratosphere and troposphere, respectively. Their primary radiative effect is to increase the scattering of sunlight away to space, either directly or in the case of the troposphere via (very uncertain) effects on clouds (refs 5, 88, and see

review in this issue by Kaufman *et al.*, pages 215–223). Thus, their primary effect is to provide a surface cooling and so a greater precipitation response compared to changes in CO₂ concentrations.

Aerosols can also produce direct tropospheric heating (refs 5, 88, and review by Kaufman), with similar effects to CO₂ on the hydrologic cycle. The effects of warming on the atmospheric moisture content, maximum rainfall amounts and the zonal-mean precipitation discussed above should all apply regardless of the cause of the warming. Aerosol-driven changes in precipitation efficiency could in principle disturb the energy-balance arguments by changing the sensible heat flux between surface and troposphere. It should be noted, however, that calculating the forcing due to well-mixed greenhouse gases, with well-defined spectral properties that can be measured in the laboratory, is much easier than the forcing due to the many possible types of aerosol, with their shorter lifetimes and complex distribution, so here we must expect rather more modelling uncertainty.

A potentially important effect of precipitation changes is on the thermohaline circulation (THC) of the North Atlantic³⁴. This brings warm water northward to give an important warming of northern mid-latitudes (refs 13, 89, and see review in this issue by Rahmstorf, pages 207–214). The apparently robust increase in high-latitude rainfall discussed above would tend to weaken the THC by stabilizing the oceanic water column in regions of deep convection (refs 13, 14, 90, and the review by Rahmstorf). The possibility has even been raised of a complete and long-term collapse in the THC, which would have a major impact on European climate in particular. Most climate models show some weakening of the THC under greenhouse-gas-induced warming, but the amount of weakening is model-dependent^{14,91}, and no constraint analogous to that shown in Fig. 2 has yet been found when comparing different models. Hence we cannot yet quantify the chance of a THC collapse during the twenty-first century.

Underestimated uncertainties

The climate modelling community has, for lack of any alternative, long used the spread of results from their AOGCMs as a subjective indicator of the range of uncertainty in climate forecasts. Modellers have, however, always understood that this range does not provide a meaningful confidence interval, as each model is designed largely as a 'best guess' and modellers share views, data, algorithms and hence, inevitably, errors⁹². The spread of results from such an 'ensemble of opportunity' is therefore likely to underestimate the true range of uncertainty in a climate forecast.

In the absence of objective probabilistic forecasts with numerical models that systematically investigate the range of possible responses of the hydrologic cycle to anthropogenic climate change^{93,94}, we have here attempted to constrain the expected changes by other means. We have, nominally, detected the influence of external forcing on recent terrestrial precipitation changes, but isolating the anthropogenic contribution is difficult as observed large-scale changes in precipitation — unlike temperature — seem to be dominated by natural forcing. Hence we cannot constrain future precipitation directly with recent precipitation trends.

Alternatively, we combine an observationally constrained estimate of 1.8–6.5 K for the 10–90% range in equilibrium warming on doubling CO₂ with a precipitation increase of 3.4% per °C of warming (plus an offset and some scatter) as obtained from a number of coupled models with idealized oceans. This gives a corresponding range for the equilibrium precipitation response of 0.6–18%. This exceeds the range of simulations of the precipitation response given by current 'best guess' climate models.

Precipitation changes predicted from climate models depend heavily on the simulations of present-day precipitation, which have many deficiencies. Hence we will continue to need conventional model development, including additional processes and exploring the impact of higher resolution. But we should not sacrifice everything on the altar of spatial resolution: ultra-high-resolution models

are too expensive to be used systematically to identify the constraints (linking observable to predicted climate variables) that are essential to probabilistic climate forecasting. We believe (see Box 1) that objective probabilistic forecasting using AOGCMs is now possible, and that it is worth rising to the challenge. A forecasting system that rules out some currently conceivable futures as unlikely could be far more useful for long-range planning than a small number of ultra-high-resolution forecasts that simply rule in some (very detailed) futures as possibilities. □

doi:10.1038/nature01092

- Palmer, T. N. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.* **63**, 71–116 (2000).
- Smith, L. A. What might we learn from climate forecasts? *Proc. Natl. Acad. Sci. USA* **99**, 2487–2492 (2002).
- Folland, C. K. *et al.* in *Climate Change 2001, The Science of Climate Change* Ch. 2 (eds Houghton, J. T. *et al.*) 101–181 (Cambridge Univ. Press, Cambridge, 2001).
- Cubasch, U. *et al.* in *Climate Change 2001, The Science of Climate Change* Ch. 9 (eds Houghton, J. T. *et al.*) 527–582 (Cambridge Univ. Press, Cambridge, 2001).
- Ramanathan, V., Crutzen, P. J., Kiehl, J. T. & Rosenfeld, D. Atmosphere—aerosols, climate and the hydrological cycle. *Science* **294**, 2119–2124 (2001).
- Carnell, R. E. & Senior, C. A. Changes in mid latitude variability due to increasing greenhouse gases and sulphate aerosols. *Clim. Dynam.* **14**, 369–383 (1998).
- Trenberth, K. E. Conceptual framework for changes of extremes of the hydrological cycle with climate change. *Clim. Change* **42**, 327–339 (1999).
- Kharin, V. V. & Zwiers, F. W. Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. *J. Clim.* **13**, 3760–3788 (2000).
- Meehl, G. A. *et al.* Trends in extreme weather and climate events: issues related to modelling extremes in projections of future climate change. *Bull. Am. Meteorol. Soc.* **81**, 427–436 (2000).
- Palmer, T. N. & Räisänen, J. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **415**, 512–514 (2002).
- Osborn, T. J., Hulme, M., Jones, P. D. & Basnett, T. A. Observed trends in the daily intensity of United Kingdom precipitation. *Int. J. Climatol.* **20**, 347–364 (2000).
- Milly, P. C. D., Wetherald, R. T., Dunne, K. A. & Delworth, T. L. Increasing risk of great floods in a changing climate. *Nature* **415**, 514–517 (2002).
- Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature* **378**, 145–149 (1995).
- Clark, P. U., Pisias, N. G., Stocker, T. F. & Weaver, A. J. The role of the thermohaline circulation in abrupt climate change. *Nature* **415**, 863–869 (2002).
- Pierrehumbert, R. T. in *Mechanisms of Global Climate Change at Millennial Time Scales* (Geophys. Monogr. 112) 339–361 (American Geophysical Union, Washington DC, 1999).
- Andronova, N. G. & Schlesinger, M. E. Causes of global temperature changes during the 19th and 20th centuries. *Geophys. Res. Lett.* **27**, 2137–2140 (2000).
- Forest, C. E., Allen, M. R., Stone, P. H. & Sokolov, A. P. Constraining uncertainties in climate models using climate change detection techniques. *Geophys. Res. Lett.* **27**, 569–572 (2000).
- Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620 (2000).
- Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117 (2002).
- Gregory, J. M., Stouffer, R., Raper, S., Rayner, N. & Stott, P. A. An observationally-based estimate of the climate sensitivity. *J. Clim.* (in the press).
- Allen, M. R. Do-it-yourself climate prediction. *Nature* **401**, 642 (1999).
- Wigley, T. M. L. & Raper, S. C. B. in *Climate and Sea Level Change: Observations, Projections and Implications* (eds Warrick, R. A., Barrow, E. M. & Wigley, T. M. L.) Ch. 7 (Cambridge Univ. Press, Cambridge, 1993).
- Sokolov, A. P. & Stone, P. H. A flexible climate model for use in integrated assessments. *Clim. Dynam.* **14**, 291–303 (1998).
- Meehl, G. A., Boer, G. J., Covey, C., Latif, M. & Stouffer, R. J. The Coupled Model Intercomparison Project (CMIP). *Bull. Am. Meteorol. Soc.* **81**, 313–318 (2000).
- Covey, C., AchutaRao, K. M., Lambert, S. J. & Taylor, K. E. Intercomparison of present and future climates simulated by coupled ocean-atmosphere GCMs. Tech. Rep. 66 (Program for Climate Model Diagnosis and Intercomparison, Livermore, 2000).
- Cubasch, U. *et al.* Monte Carlo climate change forecasts with a global coupled ocean-atmosphere model. *Clim. Dynam.* **10**, 1–19 (1994).
- Levitus, S., Antonov, I. I., Boyer, T. P. & Stephens, C. Warming of the world ocean. *Science* **287**, 2225–2229 (2000).
- Barnett, T. P., Pierce, D. W. & Schnur, R. Detection of anthropogenic climate change in the world's oceans. *Science* **292**, 270–274 (2001).
- Stott, P. A. & Kettleborough, J. A. Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726 (2002).
- Knutti, R., Stocker, T. F., Joos, F. & Plattner, G. K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**, 719–723 (2002).
- Webster, M. D. *et al.* Uncertainty analysis of global climate change projections. Tech. Rep. 73 <http://web.mit.edu/globalchange/www/rpt73.html> (Massachusetts Institute of Technology, Cambridge, MA, 2001).
- Nozawa, T., Emori, S., Nakajima, A., Takemura, T. & Numaguti, A. in *Present and Future of Modelling Global Environmental Change: Towards Integrated Modelling* (ed. Matsuno, T.) (Terra Scientific, Tokyo, 2001).
- Arrhenius, S. On the influence of carbonic acid in the air upon the temperature of the ground. *Phil. Mag. Ser. 5*, **41**, 237–276 (1896).
- Stocker, T. F. *et al.* in *Climate Change 2001: The Scientific Basis* Ch. 7 (eds Houghton, J. T. *et al.*) 417–470 (Cambridge Univ. Press, Cambridge, 2001).

35. Ingram, W. J. On the robustness of the water vapor feedback: GCM vertical resolution and formulation. *J. Clim.* **15**, 917–921 (2002).
36. Schoneveld, C. D. & Rapp, J. *Climatic Trends Atlas of Europe Based on Observations 1891–1990* (Kluwer, 1997).
37. Gaffen, D. J., Rosen, R. D., Salstein, D. A. & Boyle, J. S. Evaluation of tropospheric water vapor simulations from the atmospheric model intercomparison project. *J. Clim.* **10**, 1648–1661 (1999).
38. New, M., Hulme, M. & Jones, P. Representing twentieth-century space-time climate variability. part ii: Development of 1901–96 monthly grids of terrestrial surface climate. *J. Clim.* **13**, 2217–2238 (2000).
39. Ross, R. J., Elliott, W. P. & Seidel, D. J. Lower-tropospheric humidity-temperature relationships in radiosonde observations and atmospheric general circulation models. *J. Hydrometeorol.* **3**, 26–38 (2001).
40. Zhai, P. M. & Eskridge, R. E. Atmospheric water vapor over china. *J. Clim.* **10**, 2643–2652 (1997).
41. Elliott, W. P. On detecting long-term changes in atmospheric moisture. *Clim. Change* **31**, 349–367 (1995).
42. Bates, J. J., Jackson, D. L., Breon, F. M. & Bergen, Z. D. Variability of tropical upper tropospheric humidity 1979–1998. *J. Geophys. Res.* **106**, 32271–32281 (2001).
43. Soden, B. J., Wetherald, R. T., Stenchikov, G. L. & Robock, A. Global cooling after the eruption of Mount Pinatubo: a test of climate feedback by water vapor. *Science* **296**, 727–730 (2002).
44. Boer, G. J. Climate change and the regulation of the surface moisture and energy budgets. *Clim. Dynam.* **8**, 225–239 (1993).
45. Le Treut, H. & McAvaney, B. J. A model intercomparison of equilibrium climate change in response to CO₂ doubling. Tech. Rep. 18 (Institute Pierre Simon Laplace, Paris, 2000).
46. Schlesinger, M. E. & Mitchell, J. F. B. in *Projecting the Climatic Effects of Increasing Carbon Dioxide* (eds MacCracken, M. C. & Luther, F. M.) 81–147 (United States Dept Energy, 1985).
47. Mitchell, J. F. B., Wilson, C. A. & Cunningham, W. M. On CO₂ climate sensitivity and model dependence of results. *Q. J. R. Meteorol. Soc.* **113**, 293–322 (1987).
48. Mitchell, J. F. B. The seasonal response of a general circulation model to changes in CO₂ and sea temperatures. *Q. J. R. Meteorol. Soc.* **109**, 113–152 (1983).
49. Stott, P. A. et al. External control of twentieth century temperature by natural and anthropogenic forcings. *Science* **290**, 2133–2137 (2000).
50. Mitchell, J. F. B. et al. in *Climate Change 2001, The Science of Climate Change* Ch. 12 (eds Houghton, J. T. et al.) 697–738 (Cambridge Univ. Press, Cambridge, 2001).
51. Santer, B. D. et al. A search for human influences on the thermal structure of the atmosphere. *Nature* **382**, 39–46 (1996).
52. Tett, S. F. B., Mitchell, J. F. B., Parker, D. E. & Allen, M. R. Human influence on the atmospheric vertical temperature structure: detection and observations. *Science* **247**, 1170–1173 (1996).
53. Hulme, M., Osborne, T. J. & Johns, T. C. Precipitation sensitivity to global warming: comparison of observations with HadCM2 simulations. *Geophys. Res. Lett.* **25**, 3379–3382 (1998).
54. Sokolov, A., Forest, C. E. & Stone, P. H. A comparison of the behaviour of different AOGCMs in transient climate change experiments. Tech. Rep. 81 (MIT Joint Program on the Science and Policy of Global Change, Cambridge, MA, 2001).
55. Hulme, M., Turnpenny, J. & Jenkins, G. Climate change scenarios for the United Kingdom: the UKCIP Briefing Report (Tyndall Centre for Climate Change Research, 2002).
56. Emanuel, K. A. The dependence of hurricane intensity on climate. *Nature* **326**, 483–485 (1987).
57. Giorgi, F. et al. in *Climate Change 2001, The Science of Climate Change* Ch. 10 (eds Houghton, J. T. et al.) 585–638 (Cambridge Univ. Press, Cambridge, 2001).
58. Groisman P. Ya et al. Changes in the probability of heavy precipitation: important indicators of climatic change. *Clim. Change* **42**, 243–283 (1999).
59. Karl, T. R., Knight, R. W. & Plummer, N. Trends in high-frequency climate variability in the twentieth century. *Nature* **377**, 217–220 (1995).
60. Karl, T. R. & Knight, R. W. Secular trends of precipitation amount, frequency, and intensity in the united states. *Bull. Am. Meteorol. Soc.* **79**, 231–241 (1998).
61. Huntingford, C. et al. Regional climate model predictions of extreme rainfall for a changing climate. *Q. J. R. Meteorol. Soc.* (in the press).
62. Manabe, S. & Wetherald, R. T. On the distribution of climate change resulting from an increase in CO₂ content of the atmosphere. *J. Atmos. Sci.* **37**, 99–118 (1980).
63. Dai, A., Trenberth, K. E. & Karl, T. R. Global variations in droughts and wet spells: 1900–1995. *Geophys. Res. Lett.* **25**, 3367–3370 (1998).
64. Gregory, J. M., Mitchell, J. F. B. & Brady, A. J. Summer drought in northern mid-latitudes in a time-dependent CO₂ climate experiment. *J. Clim.* **10**, 662–686 (1997).
65. Wetherald, R. T. & Manabe, S. Detectability of summer dryness caused by greenhouse warming. *Clim. Change* **43**, 495–511 (1999).
66. Trenberth, K. E. & Hoar, T. J. The 1990–1995 El Niño/Southern Oscillation event: the strongest on record? *Geophys. Res. Lett.* **23**, 57–60 (1996).
67. Wunsch, C. The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Am. Meteorol. Soc.* **80**, 245–255 (1999).
68. Trenberth, K. E. & Hurrell, J. E. Comments on: ‘The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations’. *Bull. Am. Meteorol. Soc.* **80**, 2721–2722 (1999).
69. Timmermann, A. et al. Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature* **398**, 694–696 (1999).
70. Collins, M. Understanding uncertainties in the response of ENSO to greenhouse warming. *Geophys. Res. Lett.* **27**, 3509–3513 (2000).
71. Hulme, M. A 1951–80 global land precipitation climatology for the evaluation of general-circulation models. *Clim. Dynam.* **7**, 57–72 (1992).
72. Huffman, G. J. et al. The global precipitation climatology project (GPCP) combined precipitation dataset. *Bull. Am. Meteorol. Soc.* **78**, 5–20 (1997).
73. Spencer, R. W. Global oceanic precipitation from the msu during 1979–91 and comparisons to other climatologies. *J. Clim.* **6**, 1301–1326 (1993).
74. Xie, P. P. & Arkin, P. A. Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Am. Meteorol. Soc.* **78**, 2539–2558 (1997).
75. New, M., Todd, M., Hulme, M. & Jones, P. Precipitation measurements and trends in the twentieth century. *Int. J. Clim.* **21**, 1899 (2001).
76. Huffman, G. J. et al. Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeorol.* **2**, 36–50 (2001).
77. Trenberth, K. E. & Caron, J. M. Estimates of meridional atmosphere and ocean heat transports. *J. Clim.* **14**, 3433–3443 (2001).
78. Chen, J. Y., Carlson, B. E. & Del Genio, A. D. Evidence for strengthening of the tropical general circulation in the 1990s. *Science* **295**, 838–841 (2002).
79. Wielicki, B. A. et al. Evidence for large decadal variability in the tropical mean radiative energy budget. *Science* **295**, 841–844 (2002).
80. Trenberth, K. E. et al. Changes in tropical cloud and radiation. *Science Online* **296**, 2095a (2002).
81. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
82. Thompson, D. W. J. & Wallace, J. M. Annular modes in the extratropical circulation. *J. Clim.* **13**, 1000–1036 (2000).
83. Carnell, R. E. & Senior, C. A. An investigation into the mechanisms of changes in mid-latitude mean sea level pressure as greenhouse gases are increased. *Clim. Dynam.* **18**, 533–543 (2002).
84. Gillett, N. P. et al. How linear is the Arctic Oscillation response to greenhouse gases? *J. Geophys. Res.* **10.1029/2001JD000589** (2002).
85. Venzke, S., Allen, M. R., Sutton, R. T. & Rowell, D. P. The atmospheric response over the North Atlantic to decadal changes in sea surface temperature. *J. Clim.* **12**, 2562–2584 (1999).
86. Hoerling, M. P., Hurrell, J. W. & Xu, T. Y. Tropical origins for recent North Atlantic climate change. *Science* **292**, 90–92 (2001).
87. Corti, S. & Palmer, T. N. Sensitivity analysis of atmospheric low-frequency variability. *Q. J. R. Meteorol. Soc.* **123**, 2425–2447 (1997).
88. Ramaswamy, V. et al. in *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds Houghton, J. T. et al.) (Cambridge Univ. Press, Cambridge, 2001).
89. Seager, R. et al. Is the Gulf Stream responsible for europe's mild winters? *Q. J. R. Meteorol. Soc.* (in the press).
90. Rahmstorf, S. On the freshwater forcing and transport of the North Atlantic Thermohaline Circulation. *Clim. Dynam.* **12**, 799–811 (1996).
91. Knutti, R. & Stocker, T. F. Limited predictability of the future thermohaline circulation close to an instability threshold. *J. Clim.* **15**, 179–186 (2002).
92. Smith, L. A. in *Nonlinear Dynamics and Statistics* (ed. Mees, A. I.) 31–64 (Birkhauser, Boston, 2001).
93. Hansen, J. A., Allen, M. R., Stainforth, D. A., Heaps, A. & Stott, P. A. Casino-21: climate simulation of the 21st century. *World Resource Rev.* **13**, 187–189 (2001).
94. Stainforth, D. et al. Distributed computing for public-interest climate modeling research. *Comput. Sci. Eng.* **4**, 82–89 (2002).
95. Mitchell, J. F. B., Johns, T. C. & Senior, C. A. The transient response to increasing greenhouse gases using models with and without flux adjustment. Tech. Rep. <www.metoffice.com/research/hadleycentre/pubs/HCTN> (1998).

Acknowledgements

We are grateful to G. Hegerl and K. Trenberth for thorough reviews; to J. Gregory, C. Forest, J. Mitchell, L. Smith and I. Tracey for helpful suggestions; and to B. Booth, B. Grey and H. Lambert for help with the figures. M. Webster, M. Hulme, P. Stott and J. Gregory kindly provided unpublished data used in Figs 1, 3 and 4, and we thank C. Covey and the CMIP-2 data team for data in Figs 2, 5 and 6. M.R.A. is supported by the UK NERC and US NOAA & DoE, W.J.I. by the UK Government Meteorological Research Programme and D.A.S. by the NERC COAPEC Programme. The www.climateprediction.net programme is funded by the UK NERC, EPSRC and DTI ‘e-Science’ programmes.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Silje Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.johnson@naturejpn.com

Asia-Pacific Advertising Manager:

Keyn Johnson

naturejobs

The physics imbalance

Last week, at a meeting of the Institute of Physics in London, a team of senior women physicists addressed the question of why women are underrepresented in physics worldwide. Their resulting report, *Women Physicists Speak*, proposes a number of ways in which the balance might be redressed and careers in physics be made more flexible.

For example, women who take time off to start a family are unintentionally penalized, through a tougher path to tenure. By introducing flexible age limits for positions and providing funding sources for women who want to return to physics after a career pause, the report suggests that women physicists could make some inroads. And because there are fewer women in leadership positions, such as heads of departments or of physical-science societies, the team proposes that some of these positions could be split between women and men.

Finally, for couples who are both involved in physics, securing jobs in the same geographical location can be hard, and the physicists say that universities could help by recruiting both spouses or offering them shared academic posts.

These proposals all seem like sound starts, but perhaps restricting them to physics is too limiting. Although women are better represented in biology, more re-entry grants would be welcome for scientists thinking about leaving the bench to start a family. Schemes to help women crack the glass ceiling of management would also be helpful.

But why restrict sound quality-of-life initiatives to women? If a male physicist wants to take time out to tend to young children, or ailing parents, why not give him the option to do so without fear of risking his career? And mentoring young scientists — male or female — is something the 'seniors' could do in all disciplines.

Paul Smaglik

Naturejobs editor



Contents

CAREERS & RECRUITMENT

Chemistry meets
information technology p4

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

Chemistry plans a structural overhaul

The rising tide of data being generated by high-throughput approaches to drug screening is slowly bringing about a chemical revolution. Chemoinformatics, which marries chemistry with computer science, is becoming big business, says Eugene Russo.

As with an astronomer's fortuitous capture of a supernova in action, those armed with the right sort of telescope can witness the emergence of a new scientific discipline. First there is the increasing use of an all-inclusive buzzword — most recently something with the suffix ‘-omics’ or ‘-matics’. New programmes, journals and societies usually follow, along with rising demand for the field's practitioners. A clear definition of the field may only come later.

Those indicators are growing stronger for chemoinformatics — also called cheminformatics or chemical informatics. Just as the genomics boom created a need for bioinformaticians, an explosion in the amount of data generated by combinatorial chemistry and other high-throughput approaches to drug screening and drug design is creating a demand for chemoinformaticians. And training programmes are springing up to meet the demand.

Although the definition of chemoinformatics varies, there is always the implication of an intersection of chemistry and computer science. Helen Schofield, leader of the chemoinformatics MSc course at the University of Manchester Institute of Science

and Technology (UMIST) in Britain, calls it a blend of “chemical information” in the more traditional sense — the use of databases, searching for chemical structures, searching for keywords — with molecular modelling, computational chemistry, drug design and the optimal storage of chemical structures on computers. The university's official definition is: “The application of computational techniques to the discovery, management, interpretation and manipulation of chemical information and data extracted therefrom.”

Peter Willett, head of the University of Sheffield's MSc chemoinformatics programme, prefers to narrow the definition. He argues that the science of chemoinformatics refers specifically to the computer manipulation of two- or three-dimensional chemical structures and excludes textual information. This distinguishes the term from chemical information, largely a discipline of chemical librarians that does not include the development of computer methods.

In one of the first citations of the term, Frank Brown of healthcare-products company Johnson & Johnson based in New Brunswick, New Jersey, calls chemoinformatics a “mixing of information resources

to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug-lead identification and optimization”, (*F. K. Brown Annu. Rep. Med. Chem.* **33**, 375–384; 1998).

Working with a much more general definition, Johann Gasteiger, an organic chemist at the University of Erlangen-Nuremberg in Germany and one of the founders of the discipline, calls chemoinformatics simply

Spreading the word: postgraduate courses run by Helen Schofield (left) and Peter Willett are helping to train a new generation of chemoinformaticists.



CAREERS AND RECRUITMENT

the application of informatics to solve chemical problems. "Chemoinformatics makes the point that you're using one scientific discipline to understand another scientific discipline," says Gasteiger, who is now working on what is perhaps the first chemoinformatics textbook for graduate and undergraduate students. "It's somewhere in between, and we can only benefit if the two disciplines work together."

In a sense, the field of chemoinformatics is not really new. Scientists were searching computer databases for chemical structures 40 years ago. "I always say I did chemoinformatics for 25 years without even knowing it," says Gasteiger. In the late 1980s, he and University of Texas professor of pharmacy Robert Pearlman each independently developed software packages that convert two-dimensional molecular-structure information into three dimensions, an advance that helped to revolutionize the drug industry.

DATA EXPLOSION

Just as genome-sequencing technology has generated a data explosion in molecular biology, combinatorial chemistry and high-throughput screening have led to mountains of chemical data that must be categorized and made searchable. Drug companies, some of them struggling financially, are particularly intent on cutting costs and boosting efficiency with better ways of data sifting (see 'Eyeing pipeline efficiencies', overleaf). To a lesser but increasingly significant extent, chemical and agrochemical companies need chemoinformatics to devise more advanced products, from pesticides to polymers (see 'Chemoinformatics in chemistry', below). "Much of what is done in chemoinformatics is the natural evolution of stuff that has been going on for a very long time," says Willett.

Chemoinformatics as a field was established in the wake of its cousin discipline, bioinformatics. Wendy Warr, management consultant for chemoinformatics-related industries and long-time observer of the field, recalls some suspicion of the chemoinformatics label a few years ago. Some wondered if it had simply been created to mimic the bioinformatics label, a hot field that boasted lucrative salaries. "Some of us used to cynically think that this job title of chemoinformatics



was invented by the people doing computational chemistry or chemical information because it sounded a whole lot sexier," says Warr.

The distinction between the two fields is fuzzy. Bioinformaticians generally focus on genes and proteins, chemoinformaticians on small molecules. And both disciplines come into play when dissecting such things as protein structure and function or the binding of a small-molecule ligand to its receptor protein. "The connection between the two is getting stronger and stronger all the time," says Rich Lawson, global director of lead informatics at drug-discovery company AstraZeneca R&D in Boston.

The connection has strengthened in terms of content and analytical tools. Researchers in the pharmaceutical industry in particular collect data on much more than just small-molecule ligands. They want to know about the target proteins that they are assaying, and how those targets are related to each

Cooking up a storm: supporters believe that chemoinformatics will have a significant impact in many areas of chemistry.

Chemoinformatics in chemistry

The demand for chemoinformatics expertise has come primarily from the pharmaceutical industry. But chemistry-database creation and searching, and high-throughput technologies, have become increasingly important to the chemical industry as well.

Chemical companies share some of the analytical techniques of the drug industry, such as molecular spectroscopy, immunoassays and chromatography; some techniques, such as X-ray

fluorescence and materials testing are less relevant to pharmaceuticals. "The main focus of the drug industry is the human being as a target for the application for the products it's developing," says David Rothman, head of lead informatics at Dow Chemical based in Midland, Michigan. "In the world of basic chemistry, our targets are almost everything else." Dow, for example, might look for data on the toughness of a new polymer, its molecular weight,

tensile strength or melting point. The company is also interested in monitoring and optimizing the behaviour of chemical processes by which products are made.

But like the drug industry, basic chemical industries face information-management problems that chemoinformatics technologies could alleviate. Companies must keep track of countless chemical compounds and their properties. According to Rothman, researchers in both

the chemical and the pharmaceutical industries now spend more time digesting data than generating them, whereas years ago the reverse was true.

Rothman hired several chemoinformatics specialists last year, and hopes to hire more this year. He predicts that the ratio of basic chemistry investigators doing information technology versus those doing bench science will continue to increase. "It's only a question of how much and how fast," he says.

E.R.



Back to his roots: Martin Cadman saw a course in chemoinformatics as a way to return to science.

other. Tools such as cluster analysis, statistical approaches and neural networks are becoming increasingly relevant to both disciplines.

Both bioinformatics and chemoinformatics experts are in great demand, although the extent of that demand depends on one's definition of the field in question. Include computational chemistry under the umbrella of chemoinformatics and the job market widens considerably.

"Even now you will see quite different job descriptions from

companies claiming they want a chemoinformatics expert," says Warr. Chemoinformatics jobs can range from the management of a database of structures and biological data, to work involving combinatorial chemistry.

According to David Alker, discovery recruitment and academic liaison manager for the UK branch of the pharmaceutical company Pfizer at Sandwich in Kent, many of the current chemoinformatics-relevant jobs in Britain and Europe are in biotech start-ups. Gary Wiggins, director of programmes in bioinformatics and chemoinformatics at Indiana University in Bloomington, points out that the frequent mergers in the pharmaceutical industry require the combining of each company's huge proprietary databases, a job tailor-made for the chemoinformatician.

"It is a challenge to find people with the right skills and the right experience," says Lawson. AstraZeneca wants people adept at statistical treatment of data, data analysis, database design, information management and computer programming who also know their chemistry and even some biology. Lawson also emphasizes the desire for people with "professional skills", meaning good communicators and listeners. "Because we're looking for all that within a single person, it's a real challenge," he says.

Picking out authentic chemoinformaticians has become increasingly difficult. Often scientists who use computer technologies for chemical research call themselves chemoinformaticians, even though they can lack the core skills, according to Arnold Dion, chief executive of the US-based National Bioinformatics Institute (NBI). "They've done 100 BLAST assays that anybody can do and they suddenly reinvent themselves as a bioinformatician or chemoinformatician," he says. Formed less than a year ago, the NBI is a certification and testing service for bioinformatics and, more recently, chemoinformatics. In the case of chemoinformatics, the NBI tests things such as clustering methods, use of databases, database structures, and knowledge of UNIX and

Into the classroom

For Martin Cadman, the chemoinformatics master's degree was a way to return to science and ponder future career possibilities. After earning his chemistry degree at the University of Sheffield, Cadman headed for the UK arm of personal- and household-products firm Procter and Gamble in Newcastle upon Tyne. But the job involved more project management than scientific research. After two years, Cadman wanted to return to his chemistry roots.

As an undergraduate, he had enjoyed his limited experience with computational chemistry and molecular modelling. "I got a taste of it the first time around, but not really enough," says Cadman. He looked to a chemoinformatics programme at the University of Manchester Institute of Science and Technology in Britain as a first step towards either a PhD or

a return to industry, this time in computational chemistry. A major factor in his decision was that the degree was free. Money from the Engineering and Physical Sciences Research Council funds several UK chemoinformatics master's students each year.

After finishing a year of courses, Cadman became an intern at De Novo Pharmaceuticals in Cambridge, UK, helping to develop algorithms for drug discovery. Although the internship has enabled him to network, Cadman is not convinced that employers are sufficiently familiar with the chemoinformatics MSc. "If they're getting a graduate chemist, they know what they're getting," says Cadman. "But if it's a graduate chemoinformatics student, they didn't know the course existed in the first place. It's up to us to show them we're worth recruiting." **E.R.**

The frequent mergers in the drug industry are opening up opportunities for chemists conversant with databases, says Gary Wiggins.



Structured Query Language (SQL), which is used to maintain databases. NBI certification, awarded through participating institutes and colleges, is aimed primarily at chemists, many of whom are already employed. The course takes 20–30 weeks, roughly six hours per week.

CHEMOINFORMATICS COMES TO ACADEMIA

New degrees in chemoinformatics may also help to standardize the field. Spurred on by the needs of industry, a few programmes have been established at universities.

A select few labs, including those of Gasteiger and Pearlman, have for years been feeding the drug industry's craving for chemoinformatics. "I could probably send away each week a person to a position," says Gasteiger, referring to frequent industry enquiries about his postdocs and graduate students. Willett estimates that worldwide production of chemoinformatics specialists from academic labs amounts to fewer than 20 people each year. In the absence of formally trained chemoinformaticians,

Eyeing pipeline efficiencies

Drug companies have poured millions of dollars into new chemoinformatics and bioinformatics technologies to improve drug discovery. But so far these technologies do not seem to have significantly boosted efficiency or cut costs.

Drug-development costs have skyrocketed in recent years. A November 2001 study by the Tufts University Center for the Study of Drug Development showed that the average cost to develop a new prescription drug is US\$802 million, up from a reported \$231 million in 1987. Adjusted for inflation, the average cost

in 2000 should have been \$318 million.

Despite the rising costs, drug-development pipelines are running dry. Janice Reichert, senior research fellow at the Tufts centre, suggests that the recent profusion of mergers indicates that companies are eyeing each other's drug pipelines in the hope of improving their own.

But the Tufts study found that the increasing costs are primarily a result of the rising cost of clinical trials, and not of the money spent on chemoinformatics and related technologies. New technologies

may be having a relatively neutral role.

Johann Gasteiger, an organic chemist at the University of Erlangen-Nuremberg in Germany and a long-time developer of drug-discovery software, emphasizes that drug production is a poor measure for the success of a technology. Improvements in drug discovery will not necessarily come from new hardware and software directly, he says, but as the software user — the chemist — learns to make optimum use of the informatics tools.

Rich Lawson, global director of lead informatics at

AstraZeneca R&D in Boston, says that his company initially focused on building a core infrastructure to support developments such as combinatorial chemistry and high-throughput screening. But the company has started to shift towards mining data rather than building systems. "If we are going to invest half a million dollars in a high-content screening instrument that produces terabytes of complex data," he says, "we have to accept the fact that we also need to invest in the means to effectively manage these intellectual assets." **E.R.**

companies typically hire specialists in fields such as crystallography, and then teach them about computers and databases.

But tributaries for chemoinformatics positions may be proliferating. Programmes directed specifically towards teaching a suite of chemoinformatics skills have appeared only in the past three years. Indiana University in Bloomington, the University of Sheffield and UMIST all recently started awarding master's degrees in chemoinformatics.

In the United Kingdom, the Engineering and Physical Sciences Research Council has recognized the shortage of chemoinformatics expertise in industry, and in 1999 asked for proposals for the development of MSc programmes. Sheffield and UMIST were awarded funding. All three existing programmes have strong ties with industry. Students have the opportunity to work in industry as part of their degree requirements, and companies often provide guest lecturers. "It provides industry with a trained workforce, which it wasn't getting otherwise," says Willett.

Sheffield and UMIST are one-year programmes. The UMIST programme includes required courses on spectroscopy and crystallography, database design, bioinformatics and the use of chemical information sources such as SciFinder and CrossFire. Sheffield's programme includes courses on database design, information storage and retrieval, and object-oriented programming. Indiana University has a BS and a two-year MS in chemoinformatics. The MS has two major required areas: a course that explains commercial databases for everything from patents to spectra to reactions, and one that focuses on molecular modelling. Students can elect to take additional options such as bioinformatics.

Wiggins sees Indiana's chemoinformatics master's degree as a way to achieve a broad overview of the topic. Some students may even enter the programme with chemistry PhDs. But, says Wiggins, two years in the programme will teach a new skill set tailor-made for a company's interests.

The programmes are young, so feedback about

graduates' career paths is scarce (see 'Into the classroom', opposite). Most graduates will probably head to industry, primarily to drug companies. But others could work as chemical librarians. UMIST's Schofield suspects that programmes such as hers are beginning to get some recognition from industry; she has started to see more job advertisements that do not require PhDs, meaning that a BSc and/or MSc and the proper experience may be adequate. "That's what we're sincerely hoping, otherwise the MSc will not be worth them having," she says. "The proof will be when we see how this year's students do on the job market."

For those who cannot make it to campus in person, Indiana, Sheffield and UMIST are working together to develop distance-learning modules. People in industry would have the opportunity to earn units towards getting a degree, or to use chemoinformatics training packages that would not result in any qualification.

New programmes, informed and supported by industry, are likely to be established at other universities. Even so, academia may have a tough time establishing courses that are in tune with the changing technical demands of industry. "It's going to be three years before you set the course up," says Warr, "by which time industry has changed its mind yet again about what it might need."

Eugene Russo is a freelance science writer in Takoma Park, Maryland.

Web links

Cambridge Health Institute's Chemoinformatics glossary

♦ www.genomicglossaries.com/content/chemoinformatics_gloss.asp

Chemical Informatics at Indiana University

♦ www.indiana.edu/~cheminfo/informatics/cinform_what.html

Chemoinformatics at the University of Sheffield

♦ www.shef.ac.uk/uni/academic/I-I-M/is/courses/pggrad/mscci/mscci.html

Chemoinformatics at UMIST

♦ www.ch.umist.ac.uk/PG/msc/MScCheminf.htm

Chemoinformatics certification at the National Bioinformatics Institute

♦ www.bioinformatics.com/chemoinfo.htm



Rich Lawson sees a growing connection between bioinformatics and its chemical counterpart.